(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification⁷: **C12Q 1/68**

(21) International Application Number: PCT/US01/11609

(22) International Filing Date: 10 April 2001 (10.04.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/196,063   10 April 2000 (10.04.2000)   US
60/196,258   11 April 2000 (11.04.2000)   US

(71) Applicant and
(72) Inventor: ASHBY, Matthew [US/US]; 91 Longfellow Road, Mill Valley, CA 94941 (US).

(74) Agents: HALEY, James, F., Jr. et al.; c/o Fish & Neave, 1251 Avenue of the Americas, New York, NY 10020 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:
— as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii)) for all designations
— of inventorship (Rule 4.17(iv)) for US only

(54) Title: METHODS FOR THE SURVEY AND GENETIC ANALYSIS OF POPULATIONS

16s rRNA Polymorphic Regions
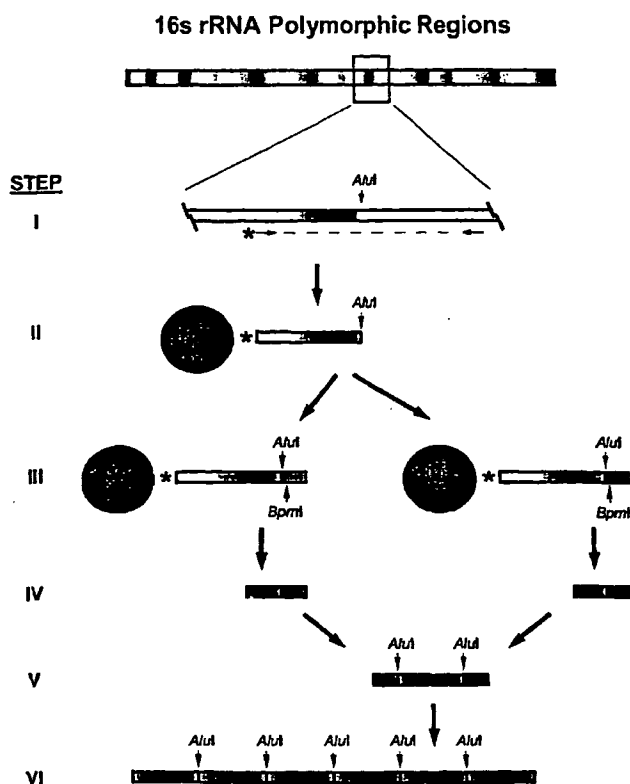
(57) Abstract: The present invention relates to methods for performing surveys of the genetic diversity of a population. The invention also relates to methods for performing genetic analysis of a population. The invention further relates to methods for the creation of databases comprising the survey information and the databases created by these methods. The invention also relates to methods for analyzing the information to correlate the presence of nucleic acid markers with desired parameters in a sample. These methods have application in the fields of geochemical exploration, agriculture, bioremediation, environmental analysis, clinical microbiology, forensic science and medicine.

WO 01/77392 A2

METHODS FOR THE SURVEY AND
GENETIC ANALYSIS OF POPULATIONS

This application claims the benefit of United States Provisional

Application 60/196,063, filed April 10, 2000 and United States Provisional

5   Application 60/196,258, filed April 11, 2000.

TECHNICAL FIELD OF THE INVENTION

The present invention relates to methods for performing surveys of

the genetic diversity of a population. The invention also relates to methods for

10   performing genetic analyses of a population. The invention further relates to

methods for the creation of databases comprising the survey information and the

databases created by these methods. The invention also relates to methods for

analyzing the information to correlate the presence of nucleic acid markers with

desired parameters in a sample. These methods have application in the fields of

15   geochemical exploration, agriculture, bioremediation, environmental analysis,

clinical microbiology, forensic science and medicine.

BACKGROUND OF THE INVENTION

Microbes have been used previously as biosensors to identify

chemicals in the environment. For instance, microbes have been utilized as

20   biosensors for the presence of nitrates (Larsen, L. H. et al., 1997, A microscale

NO_3^- biosensor for environmental applications. *Anal. Chem.* 69:3527-3531),

metals (Virta M. et al., 1998, Bioluminescence-based metal detectors. *Methods*

Mol. Biol. 102:219-229), and a variety of hydrocarbons (Sticher P. et al., 1997,

Development and characterization of a whole-cell bioluminescent sensor for

25   bioavailable middle-chain alkanes in contaminated groundwater samples. *Appl.*

Environ. Microbiol. 63(10):4053-4060). In these examples however, the indicator

- 2 -

microbes are not native species, but rather, the product of recombinant
manipulations designed for specific applications. These modifications involve
coupling the nutrient sensing machinery of well-characterized bacterial strains with
reporter genes to aid identification. This approach is limited, however, by the
5   metabolic diversity of a few well-characterized bacterial strains. In contrast, the
large and diverse pool of microbes in the environment represents a source of
biosensors for a much larger range of applications than currently exists. Thus, there
is a need to identify and use other microbes, especially those found *in situ*, as
biosensors.

10   Microbes also have an important impact on health and medicine.
Estimates have been made there may ten times the number of microbial cells
associated with the human body as there are human cells. Many microbial cell
populations that are associated with the human body play a beneficial role in
maintaining health. For instance, gut microflora is important for proper digestion
15   and absorption of nutrients and for production of certain factors, including some
vitamins. In general, the human immune system is able to keep the bacterial
populations of the human body in check and prevent the overgrowth of beneficial
microbial populations and infection by detrimental microbial populations.
Nevertheless, the list of human diseases that are now attributed to microbial
20   pathogens is growing. However, nearly all of the information regarding the
relationships between microbes and human disease have been gained from
approaches that require culture of microbial species.

            Two examples of diseases where the causative agents were identified
through molecular methods include bacillary angiomatosis (Relman, D.A. et al.,
25   1990, New Engl. J. Med. 323: 1573) and Whipple's disease (Wilson, K.H. et al.,
1991, Lancet 338: 474). Further, the central aspects of atherosclerosis are
consistent with the inflammation that results from infection. DNA sequences from

- 3 -

Chlamydia have been identified from atherosclerotic lesions and has led to suggestions that this organism plays a role in the disease.

In addition, bacterial infections have become an increasing health problem because of the advent of antibiotic-resistant strains of bacteria. Further,

5   microbial infections caused by bacteria or fungi that do not usually infect humans may be a problem in immunocompromised individuals. Further, individuals in developing countries who may be malnourished or lack adequate sanitary facilities may also support a large load of opportunistic bacteria, many of which may cause sickness and disease. In veterinary medicine, livestock living in close quarters also

10   may be prey to infections caused by a variety of different types of microbes. Thus, there is a need to develop sensitive methods of identifying many different types of microbes without having to cultivate them first in order to treat or prevent microbial infections in humans and other animals.

Assays for microbial contamination is an important component of

15   food testing as well. A large number of different types of microbes may contaminate food for humans or animals. Thus, an ability to test food for contamination quickly and effectively is critical for maintaining food safety. However, many of the microbes responsible for causing sickness in humans and animals are difficult to isolate or identify.

20   Assays for microbial populations also has use in fields such as forensic science. Over the past ten to twenty years, scientists have determined that microbial populations change when bodies begin to decay, and have begun to identify certain microbial species that are indicative of decomposition (Lawrence Osborne, Crime-Scene Forensics; Dead Men Talking, New York Times, December

25   3, 2000). However, only a few microbial species that may be useful in these analyses has been identified.

The problem of determining genetic diversity is not confined to microbial populations. Antibody diversity is critical for a proper immune response.

- 4 -

During B cell differentiation, antibody diversity is generated in the heavy and light chains of the immunoglobulin by mechanisms including multiple germ line variable (V) genes, recombination of V gene segments with joining (J) gene segments (V-J recombination) and recombination of V gene segments with D gene segments and J

5   gene segments (V-D-J recombination) as well as recombinational inaccuracies. Furthermore, somatic point mutations that occur during the lifetime of the individual also lead to antibody diversity. Thus, a huge number of different antibody genes coding for antibodies with exquisite specificity can be generated. T cell receptor (TCR) diversity is generated in a similar fashion through recombination between

10   numerous V, D and J segments and recombinational inaccuracies. It has been estimated that 10^{14} Vδ chains, more than 10^{13} β chains and more than 10^{12} forms of Vα chains can be made (Roitt, I. et al., Immunology, 3rd Ed., 1993, pages 5.1-5.14). A knowledge of the antibody or TCR diversity in a particular individual would be useful for diagnosis of disease, such as autoimmune disease, or for

15   potential treatment.

The identification of microbes, especially soil microbes, has traditionally relied upon culture-dependent methods, whereby the detection of a microbial species depends upon the ability to find laboratory conditions that support its growth. To this end, 96-well plates have been commercially developed to

20   identify microbes with different metabolic requirements. For instance, BioLog plates incorporate 96 different media formulations into the wells of a 96-well plate. Despite these efforts, it is now accepted that far fewer than 1% of microbes can propagate under laboratory conditions (Amann, R. I. et al., 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation.

25   *Microbiol. Rev.* 59:143-169).

The widespread interest in genomics has created many exciting new technologies for the parallel quantitation of thousands of distinct nucleic acid sequences simultaneously. While still in their infancy, these technologies have

- 5 -

provided unprecedented insight into biology. To date, these technologies have predominately been utilized in pharmaceutical and agricultural applications. Genome expression profiling has gained general acceptance in biology and is likely to become commonplace in all academic, biotechnology and pharmaceutical

5   institutions in the 21st century. For instance, Serial Analysis of Gene Expression (SAGE) is a hybridization-independent method designed to quantitate changes in gene expression (Velculescu, V. E. et al., 1995, Serial analysis of gene expression. *Science* 270:484-487 and United States Patent 5,866,330). However, SAGE only measures RNA levels from tissues or organisms, and is not suitable for examining

10   genetic diversity.

The widespread interest in genomics has also led to the development of many technologies for the rapid analysis of tens of thousands of nucleic acid sequences. One such technology is the DNA chip. Although this approach had been used as a diagnostic for distinguishing between several species of the genus

15   *Mycobacterium* (Troesch, A., et al., 1999, *Mycobacterium* species identification and rifampin resistance testing with high-density DNA probe arrays. *J. Clin. Microbiol.* 37:49-55), it has limited utility for an environmental microbial survey for two reasons. First, the sequence of the target DNAs to be analyzed must be known in order to synthesize the complementary probes on the chip. However, the vast

20   majority of environmental microbes have not been characterized. Second, DNA chips rely on hybridization of nucleic acids which is subject to cross hybridization from DNA molecules with similar sequence. However, the resolving power of a hybridization-based approach is limited because one must identify regions of DNA that do not cross-hybridize, which may be difficult for related microbial species.

25   Genomic technologies and bioinformatics hold much untapped potential for application in other areas of biology, especially in the field of microbiology. However, to date there has not been a method to rapidly and easily determine the genomic diversity of a population, such as a microbial or viral

- 6 -

population. Further, there has not been a method to easily determine the antibody or TCR diversity of a population of B or T cells, respectively. Thus, there remains a need to develop such methods in these areas.

5   BRIEF SUMMARY OF THE INVENTION

The present invention solves this problem by providing methods for rapidly determining the diversity of a microbial or viral population and for determining the antibody or TCR diversity of a population of B or T cells. The present invention relies on hybridization-independent genomic technology to quickly

10   "capture" a portion of a designated polymorphic region from a given DNA molecule present in a population of organisms or cells. This portion of the DNA molecule, a "marker," is characteristic of a particular genome in the population of interest. The marker can be easily manipulated by standard molecular biological techniques and sequenced. The sequence of a multitude of markers provides a measure of the

15   diversity and/or identity of a population. In one aspect, the invention provides a method, Serial Analysis of Ribosomal DNA (SARD), that can be used to distinguish different members of a microbial population of interest.

In another aspect, the invention provides a method for analyzing a designated polymorphic region from a population of related viruses using method

20   steps similar to those described for SARD. In a further aspect, the invention provides a method for analyzing the variable regions from the immunoglobulins or TCR genes of a population of immune cells using methods steps similar to those described for SARD.

In another aspect of the invention, a method is provided for

25   analyzing a population based upon an array of the masses of peptides that are encoded by polymorphic sequences of particular DNA molecules in a region of interest. In a preferred embodiment, the region of interest is a designated

- 7 -

polymorphic region from an rDNA gene from each member of a microbial population.

In another aspect of the invention, a method is provided for analyzing the information provided by the above-described methods. The method

5   enables the creation of a diversity profile for a given population. A collection of diversity profiles provides an accurate representation of the members present in a population. These diversity profiles can be entered into a database along with other information about the population. The diversity profiles can be used with various correlation analyses to identify individual, or sets of individuals that correlate with

10   each other. The correlation analyses can be used for diagnostic or other purposes. In another aspect, the invention provides databases comprising various diversity profiles. In a preferred embodiment, the diversity profile is obtained by SARD.

In yet another aspect of the invention, a method is provided for identifying a diversity profile, as described above, that correlates with a parameter

15   of interest. In a preferred embodiment, the diversity profile is a profile of the microbial populations that correlate with the presence of mineral deposits and/or petroleum reserves. In another preferred embodiment, the diversity profile is a profile of populations of different antibodies or TCR that correlate with a specific disease state, such as an autoimmune disorder.

20   In a still further aspect, the invention provides a method for locating mineral deposits or petroleum reserves comprising identifying one or more nucleic acid markers that correlate with the presence or mineral deposits or petroleum reserves, isolating nucleic acid molecules from an environmental sample, determining whether the nucleic acid markers are present in the environmental

25   sample, wherein if the nucleic acid markers are present, then the area from which the environmental sample was obtained is likely to have mineral deposits or petroleum reserves.

- 8 -

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows a representation of a 16S rDNA gene from Bacteria illustrating polymorphic regions (shown as dark bands) and constant regions (shown as light bands).

Figure 2 shows a schematic representation of the SARD method for isolating polymorphic sequence tags from Bacterial rDNA.

Figure 3 shows a number of representative members of the domain Bacteria with their taxonomic relationships.

Figure 4 shows a number of representative members of the domain Archaea with their taxonomic relationships.

Figure 5 shows examples of Marker-Marker correlation scatter plots. Each point represents a single sample population.

Figure 6 shows examples of various Marker-Parameter scatter plots. Each point represents a single sample population.

Figures 7A, 7B and 7C shows scatter plots comparing Marker Diversity Profile (MDP) profiles. Each point represents one marker. Fig. 7A: Significant correlation with all markers within the MDP. Fig. 7B: No correlation found using all the markers within the MDP. Fig. 7C: Same plot as B except that significant correlation is found using a subset of markers.

Figure 8 shows a schematic for generating a marker diversity profile matrix database. The first step 100 involves assigning N an integer value of 1 corresponding to the first MDP data structure. The second step 105 involves assembling polymorphic markers from a sample. Examples of methods for assembling such markers are described in this application and include, but are not limited to, SARD (Figure 2, Tables I and II) or mass tag compilation (Table III). The next step 110 involves detecting the polymorphic markers. Examples of methods to detect polymorphic markers described above include DNA sequence analysis of SARD tags and MALDI-TOF analysis of mass tags, respectively. This

- 9 -

detection step includes detecting the presence and abundance of each marker in the sample. The next step 115 involves the conversion or transduction of the MDP into an electrical signal output. Generally, this process is a linear electronic conversion of the data into a digital signal. Step 120 involves detecting parameters that are

5   associated with the sample N. Step 125 involves transducing the sample parameter data, which may include, without limitation, such parameters as pH, grain size, elemental analysis and/or organic analysis, into an electrical signal output in the form of a digital signal. Step 130 involves storing into the memory of a computer the output signal from each MDP into a matrix data structure and associating it with

10   sample parameters. The next step is a decision block 135 where if all the data structures have not been completed the routine advances to step 140 where n is incremented and the data structure generating steps 105-130 are repeated for the n +1th marker diversity profile. Once all the data structures are completed the routine proceed to step 145 to form the MDP database from n data structures. Each

15   marker is assigned a unique identifier along with its relative abundance in the population. This information is also optionally indexed with other known parameters that are associated with the sample including, for instance, the time, date, elevation and geographical location. These signals are digitized and stored in the memory of a computer.

20   Figure 9 shows a schematic of the steps involved in determining parameters associated with an MDP. A marker diversity profile 200 is created for a sample. The marker diversity profile is subject to a comparison function 205 which compares the profile with resident marker diversity profiles in the database. Step 210 is a decision block where a query is made whether the new marker diversity

25   profile equals a resident marker diversity profile in the database. If the query returns a "Yes" the new marker diversity profile is deduced 215 to share the same parameters as the resident marker diversity profile. Since the parameters associated with the resident marker diversity profile are characterized, the parameters associate

- 10 -

with the new marker diversity profile are identified.

If the query in decision block **210** returns a "No" the routine proceeds to decision block **220** which queries whether the new marker diversity profile is a subset of a resident profile in the database. If the query returns a "No"

5   the parameters remain undefined **225**. If the query returns a "Yes" the routine proceeds to step **230**. Step **230** is an optional step to determine the correlation between members of the common subset of markers and may either be performed for each new profile or may be queried from a matrix table of pre-calculated values from existing profiles. Such values generally would be maintained in a relational

10 database. If this step is not performed all common markers are parsed into groups of individual markers and treated as correlated groups **255**. If the marker-marker correlation is performed between the common subset of markers, the routine proceeds to decision block **235** which queries whether all of the common markers are correlated. If the query returns a "Yes" the markers are correlated with the

15 parameters **240** resident in the database. If none of the markers are correlated with a parameter, the parameter(s) remain undefined **245** whereas if the markers are correlated with a parameter, the parameter is deduced to be associated with the marker diversity profile **250**. If the decision block query **235** returns a "No", the common markers are sorted into groups of correlated markers **255**. The first

20 correlated marker group N **260** is subject to a decision block **265** that queries whether the markers in this group are correlated with a parameter. A "No" determines that the parameters remain undefined. If the marker(s) are correlated with a parameter, the parameter is deduced to be associated with the marker diversity profile. Steps **260-275** are repeated in steps **280-295** for each correlated

25 group of markers. The groups of correlated markers may be comprised of a single or multiple markers. The confidence level in deducing that a parameter is associated with a marker diversity profile is determined by the level of correlation between the marker(s) and the parameter. Therefore, sets of correlated markers are expected to

- 11 -

be more robust indicators of any given parameter.

Figure 10 shows a schematic for generating a marker diversity profile matrix database. The first step **300** involves assigning N an integer value of 1 corresponding to the first MDP data structure. The second step **305** involves assembling polymorphic markers from a sample. Examples of methods for assembling such markers are described in this application and include, but are not limited to, SARD (Figure 2, Tables I and II) or mass tag compilation (Table III). The next step **310** involves detecting the polymorphic markers. Examples of methods to detect polymorphic markers described above include DNA sequence analysis of SARD tags and MALDI-TOF analysis of mass tags, respectively. The next step **315** involves the conversion or transduction of the MDP into an electrical signal output. Generally, this process is a linear electronic conversion of the data into a digital signal. Step **320** involves storing into the memory of a computer the output signal from each MDP into a matrix data structure associating each MDP with a geographic coordinates such as longitude and latitude. The next step is a decision block **325** where if all the data structures have not been completed the routine advances to step **330** where n is incremented and the data structure generating steps **305-320** are repeated for the n +1th marker diversity profile. Once all the data structures are completed the routine proceed to step **335** to form the MDP database from n data structures. Each marker is assigned a unique identifier and indexed with its relative abundance in the population. These signals are digitized and stored in the memory of a computer.

Figure 11 shows a schematic for mapping applications using marker diversity profiles. Marker diversity profile data **400** can be processed in several ways to create maps that provides significant environmental information. In one example **405**, each marker diversity profile in a database can be correlated with every other marker diversity profile in a pairwise manner to create a correlation matrix. By appending this data to the geographical coordinates of each sample **410**,

- 12 -

a map can be constructed that reflects the correlation values of physically neighboring sample sites. Preferably the correlation values will be color coded to reflect the level of correlation. The color is chosen from a reference color spectrum that is indexed to correlation values between 0-1.

5   Marker diversity profiles **400** can also be processed into maps at the individual marker or correlated marker group level. This approach is preferable since subsets of markers are likely to correlate to fewer number of sample associated parameters. Each marker in a marker diversity profile database is correlated with every other marker in the database in a pairwise manner to create a

10   correlation matrix **415**. The source database can either be composed of marker diversity profiles from a single geographic area or several distinct areas. In step **420**, the markers from one geographic area are sorted into groups based upon their level of correlation. In step **425**, the relative representation of the correlated marker group N is determined along with its geographical coordinates for each marker

15   diversity profile in a geographic area. A map is constructed **430** where the relative abundance of each correlated marker group is color-coded with its geographical coordinates. Steps **425** and **430** are repeated as in **435** and **440** for each correlated group of markers.

          Fig. 12 shows oligonucleotides useful for amplifying nucleic acid

20   molecules for SARD.

          Fig. 13 shows the use of the SARD strategy for Eubacteria. The double-underlined sequence and the wavy-underlined sequence represent the sequence tags for the two pools and the single-underlined sequence delineates the *Bpm*I recognition site.

25   Fig. 14 is a graphical representation of a SARD analysis of a defined population.

          Fig. 15 shows the sequence of SARD tags identified from Wy-1 sample. The number is parentheses indicates the number of tags having that

- 13 -

sequence.

Fig. 16 shows SARD tags identified from Wy-2 sample. The number is parentheses indicates the number of tags having that sequence.

Fig. 17 is a graphical representation of the number and abundance of
5   SARD tags. The upper panel shows the Wy-1 SARD Tag Diversity Profile and the lower panel shows the Wy-2 SARD Tag Diversity Profile.

DETAILED DESCRIPTION OF THE INVENTION

The extent of the diversity of microbes in our environment has only recently been recognized. With the advent of the polymerase chain reaction (PCR)
10   and small subunit ribosomal DNA (rDNA) sequence analysis, researchers have been able to detect and perform phylogenetic analyses on individual microbes without first cultivating the microbes of interest. This molecular phylogenetic approach has significantly changed our view of microbial evolution and diversity (Woese, C. R., 1987, Bacterial evolution. *Microbiol Rev.* 51(2):221-71; Pace, N. R., 1997; A
15   molecular view of microbial diversity and the biosphere. *Science.* 276(5313):734-40). For instance, the earliest life forms are now thought to have utilized inorganic compounds for nutrition rather than compounds based upon organic carbon. In addition, the vast proportion of biological diversity is now known to be due to microbial species. Estimates have been made that there may be more than ten
20   thousand *distinct* species of microbes in a single gram of soil. Figures 3 and 4 show some of the representative members of the domains Bacteria and Archaea, respectively, that may be found in environmental samples.

Microbes inhabit virtually all niches including extreme environments with temperatures between 20° and 250° F. Microbes have even been isolated from
25   deep petroleum reservoirs more than a mile beneath the earth's surface (Jeanthon, C. et al., 1995, *Thermotoga subterranea* sp. nov., a new thermophilic bacterium isolated from a continental oil reservoir. *Arch. Microbiol.* 164:91-97). In order to

- 14 -

prevail under such diverse conditions, microbes have made remarkable adaptations and have attained the ability to utilize unusual carbon and mineral resources that are immediately available. These physiological and metabolic adaptations that enable some microbes to inhabit a particular niche may also restrict their distribution to

5   such areas. Numerous examples of environmental parameters that lead to restrictions of microbial distribution are well known and are usually dictated by a species' specific metabolic program (e.g. obligate nature of the carbon, nitrogen and energy source).

        Microbes that have highly defined nutrient requirements are likely to

10 have a restricted distribution in the environment. Thus, the microbes' dependence on the presence of a particular resource to proliferate can serve as the basis for an assay to identify the presence, and characterize the distribution, of various features in the environment, such as biological, chemical and geochemical features. In other words, microbes can function as environmental biosensors.

15   In one aspect of this invention, the ability of microbes to function as environmental biosensors is used to identify particular environmental states. In a preferred embodiment, a profile of a microbial population is used to identify one or more parameters of a particular environmental state. In a more preferred embodiment, a microbial population profile is used to identify areas that are likely to

20 have mineral deposits and/or petroleum reserves. In another preferred embodiment, a microbial population profile is used in forensic science to identify decomposition of a body or to associate an individual to another individual, to an object or to a location. In yet another preferred embodiment, a microbial population profile is used to identify microbial contamination of human and animal foodstocks. In yet

25 another preferred embodiment, the profile is used to diagnose human or animal disease.

- 15 -

Definitions

Unless otherwise defined, all technical and scientific terms used herein have the meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. The practice of the present invention employs,

5   unless otherwise indicated, conventional techniques of chemistry, molecular biology, microbiology, recombinant DNA, genetics and immunology. See, e.g., Maniatis et al., 1982; Sambrook et al., 1989; Ausubel et al., 1992; Glover, 1985; Anand, 1992; Guthrie and Fink, 1991 (which are incorporated herein by reference).

A microbe is defined as any microorganism that is of the domains

10   *Bacteria, Eukarya* or *Archaea*. Microbes include, without limitation, bacteria, fungi, nematodes, protozoans, archaebacteria, algae, dinoflagellates, molds, bacteriophages, mycoplasma, viruses and viroids.

A marker is a DNA sequence that can be used to distinguish or identify a particular gene, genome or organism from another. In one embodiment, a

15   marker may be generated by one of the methods described herein. A marker represents one or a limited number of taxonomic species or genes. In a preferred embodiment, a marker represents a single taxonomic species or gene. In one embodiment, the marker represents a single microbial species. In another embodiment, the marker represents a single viral species or type. In another

20   embodiment, the marker represents a single immunoglobulin or TCR variable domain.

A marker diversity profile (MDP) is a data set that is obtained from each population sample and that contains a collection of markers. In a preferred embodiment, the MDP also comprises other information, including all known

25   parameters associated with a particular population sample. Such parameters that relate to environmental samples may include inorganic components (obtained through atomic adsorption analysis), organic components (obtained through GC-MS or LC-MS), grain size analysis, pH, and salinity. Parameters that relate to

- 16 -

medical samples would include, but are not limited to, a complete medical history of the donor. In a preferred embodiment, the markers are obtained by SARD.

Methods for the Genetic Analysis of Populations

Ribosomes, which are comprised of numerous ribosomal proteins

5   and three ribosomal RNA (rRNA) molecules, are a key component of protein synthesis. The 16S subunit rRNA, which is encoded by the 16S rDNA gene, has been the focus of much attention in microbial phylogenetic studies. The 16S rDNA sequence is highly conserved between taxonomic groups, yet also possesses regions that are highly polymorphic (Figure 1). Moreover, the rate of change in the RNA

10   sequence is thought to have been relatively constant over evolutionary time, enabling scientists to determine the relative relatedness of different organisms.

Typical molecular microbial analyses involve utilizing the highly conserved regions of the 16S rDNA to amplify the roughly 1,500 bp gene. The sequence of the PCR-amplified product is determined and compared with other

15   known rDNA sequences. Although this approach is highly informative, it is not amenable to a rapid survey of an environmental microbial community.

The instant invention provides methods for quickly and easily calculating the genetic diversity of a population. The methods uses hybridization-independent genomic technologies to overcome the previously-identified problems

20   of determining genetic diversity. This method may be used for any population of cells, viruses or organisms which comprise at least one DNA molecule that comprises regions of high sequence conservation interspersed with polymorphic sequences, wherein the polymorphic sequences can be used to distinguish different members of the population of interest. One aspect of the present invention

25   describes a method (SARD) that can capture a designated polymorphic region from a given DNA molecule present in the members of a microbial community. In a preferred embodiment, the DNA molecule is a 16S rDNA molecule. In another

- 17 -

embodiment, the DNA molecule is the intergenic region between the 16S and 23S rDNA genes. In another embodiment, the method is used to identify the genetic diversity of a population of viral samples or of cells or organisms infected with a population of viruses. In another embodiment, the method is used to identify the

5   diversity of immunoglobulin and/TAR genes in a population of B and/or T cells.

The method may be performed as follows (see Figure 2):

Step I. Sample Preparation and DNA Amplification by PCR

Samples may be obtained from any organism or region desired. For environmental microbial analyses, samples may be obtained from, without limitation,

10   buildings, roadways, soil, rock, plants, animals, cell or tissue culture, organic debris, air or water. For medical microbial analyses, samples may be obtained from, without limitation, humans, animals, parasites, water, soil, air and foodstuffs. For viral analyses, samples may be obtained from, without limitation, viral culture ' stocks, humans, animals, plants, cell or tissue culture and microbes. For

15   immunoglobulin or TCR analyses, samples may be obtained from, without limitation, humans, animals or cell or tissue cultures. DNA molecules from the sample of interest may be isolated by any method known in the art. See, e.g., Sambrook et al., 1989 and Ausubel et al.., 1992. In a preferred embodiment, DNA is obtained as described by Yeates et al., "Methods for Microbiological DNA

20   Extraction from Soil for PCR Amplification," Biological Procedures Online, Volume 1, May 14, 1998, available at www.science.uwaterloo.ca/bpo; Liu et al., Applied and Environmental Microbiology (1997) 63: 4516-4522; and Tsai et al., Applied and Environmental Microbiology (1992) 58: 2292-2295. The DNA molecules do not have to be completely purified but only need be isolated to the

25   point at which PCR may be performed.

Environmental microbes often exist in biofilms (Costerton, J.W., et al., 1999, Bacterial biofilms: a common cause of persistent infections. *Science*

- 18 -

284(5418):1318-1322) or in tight association with solid surfaces. Microbial DNA
from a sample of interest is isolated by one of several methods that are widely
known to those skilled in the art and are described in the literature (Gillan, D.C. et
al., 1998, Genetic diversity of the biofilm covering *Montacuta ferruginosa*

5   (Mollusca, bivalvia) as evaluated by denaturing gradient gel electrophoresis analysis
and cloning of PCR-amplified gene fragments coding for 16S rRNA. *Appl.
Environ. Microbiol.* 64(9):3464-72).

The samples may be selectively enriched before they are isolated by
any method known in the art. In one embodiment, for a population of microbes that

10   are known or suspected to feed on a hydrocarbon source such as propane, the
hydrocarbon may be added to the environment in which the microbes live for a
period of time before the microbes are harvested. In another embodiment, a viral
population may be cultivated in cells before they are isolated. In a further
embodiment, B and T cells may be expanded in culture before isolation. It may be

15   easier to obtain sufficient sample amounts if a population is expanded before
isolation. However, this must be weighed against the possibility that expansion will
alter the ratio of different members of the population to each other.

In general, the primers used for amplification are designed to
hybridize to a region of the DNA that is highly conserved between members of the

20   population. Further, the primers should flank a polymorphic region the partial
sequence of which should provide diagnostic information regarding the genetic
diversity of the population. For instance, for the 16S rDNA gene, primers are
designed to hybridize to a highly conserved region of the 16S rDNA gene flanking a
polymorphic region (see Figure 2). In an immunoglobulin gene, the primers are

25   designed to hybridize to a region of the B cell DNA flanking the V-J recombination
site. Alternatively, primers may be designed that bind to the relatively constant
regions within certain regions of the V-J gene that flank a polymorphic region. See
Roitt et al., Immunology, 3rd Ed., 1993, pp. 5.2-5.14, herein incorporated by

- 19 -

reference, which shows regions of variability and conservation within immunoglobulin and TCR genes. One having ordinary skill in the art following the teachings of the specification will recognize that other genes that have regions that are highly conserved between members of the population and that flank

5   polymorphic regions may be used in the design of primers.

The primers should also be designed to flank a region of DNA that comprises a restriction site for a restriction enzyme. In a preferred embodiment, the restriction enzyme is a cuts at a four-basepair recognition site, such as *Alu*I (see Figure 2). Furthermore, the restriction site should be near but not in the

10   polymorphic region of the gene of interest. In a preferred embodiment, the restriction site should be one that is present in the gene of interest in a majority of the known species or genes.

A single set of primers may be used or multiple sets of primers may be used. A single set of primers may be used if it is known that the region of DNA

15   to which the primers will bind is very highly conserved. Alternatively, if it is known that there is some variation in the conserved region, multiple sets of primers may be used to bind to the conserved DNA region. Using multiple sets of primers may be useful to identify more members of a population, especially those members of the population that exhibit less sequence identity in the conserved areas of a nucleic acid

20   sequence. In one embodiment, four to ten sets of primers may be used to identify members of a population. Alternatively, the primers used may be degenerate, such that different molecules within a primer population will include a different base at one or more specific sites in the primer. For instance, a primer may have a site that has either cytosine or thymidine. The purpose of making primers degenerate is to

25   increase the number of different DNA molecules that will hybridize to a particular primer. Methods of making degenerate primers are well-known in the art.

The primers used in this and in subsequent steps are generally of a length sufficient to promote specific hybridization to a DNA sequence. The primers

- 20 -

generally have a length of at least 12 bases, more preferably at least 15 bases, even

more preferably at least 18 bases. The primers may have a length of up to 60 bases,

although usually most are under 40 bases in length. Primers may include both bases

that are naturally found in DNA, such as adenine, guanine, cytosine and thymidine,

5   and may also include nucleotides that are not usually found in DNA, such as inosine.

                One of the primers (the "upstream" primer) should be modified to

incorporate a moiety that can be used to bind the PCR product to a solid support.

The upstream primer is defined as the primer that is located on the opposite side of

the polymorphic region of interest relative to the flanking four-base restriction site.

10   A number of different binding moieties are known in the art. In a preferred

embodiment, the moiety is biotin. In another preferred embodiment, the moiety is

digoxigenin or six histidines.

                PCR is performed using the primers to amplify a subregion that

contains a polymorphic site of interest. Methods for performing PCR are well-

15   known in the art. In one embodiment, the PCR products are normalized or

subtracted by methods known in the art to lower the representation of the dominant

sequences. Exemplary methods are described in Sambrook et al., 1989, Ausubel et

al., 1992; Glover, 1985; Anand, 1992 (which are incorporated herein by reference).

Step II. Digestion of the Amplified Fragment and Binding to Solid Support

20   The amplified fragment is cut with the restriction enzyme as

discussed in Step I. Any restriction enzyme may be used in this step so long as it is

cuts at a site immediately adjacent to the polymorphic sequence. In a preferred

embodiment, the enzyme is a four-base restriction enzyme. Examples of four-base

restriction enzymes are well-known in the art and include many that are

25   commercially available. See, e.g., New England Biolabs Catalog 2000, herein

incorporated by reference. Examples of four-base restriction enzymes include,

without limitation, *AluI, Bsh*1236I, *DpnI, HpaII, MboI, MspI, PalI* (an

- 21 -

isoschizomer of *Hae*III), *Rsa*I, *Sau*3AI and *Taq*I. After restriction, the DNA

fragment is bound to a solid support. Numerous solid supports to immobilize DNA

are known in the art. Examples include, without limitation, streptavidin beads,

which would bind to a PCR product labeled with biotin, and anti-digoxigenin beads,

5   which would bind to a PCR product labeled with digoxigenin, and beads conjugated

to nickel, which would bind to a six-histidine labeled product. In a preferred

embodiment, streptavidin beads are used (Figure 2).

Since the SARD tag position is dictated by the first restriction

enzyme recognition site distal to the biotinylated primer used in the initial PCR

10   reaction, there may be cases in which the first restriction enzyme recognition site is

located within a conserved region of the gene of interest. In general, this will not be

a problem because even though the tags from the conserved region may not be

informative, most tags derived by SARD will be from a polymorphic region and will

be informative. However, if one desires to decrease the number of tags that contain

15   information from a conserved region of a gene rather than from a polymorphic

region, one may purify the desired PCR products after restriction. In a preferred

embodiment, one may do this by gel purifying those PCR products that have the

expected size.

Step III. Digestion of the Amplification Product and Ligation to Linkers

20   The immobilized products are split into two pools and linkers are

attached to the immobilized products of each pool. Each linker is a double-stranded

synthetic DNA molecule comprising a specific DNA sequence. Both linkers

incorporate a Type IIS restriction enzyme site. In a preferred embodiment, the two

linkers incorporate the same Type IIS restriction enzyme site. Each of the two

25   linkers also comprises a DNA sequence that specifically hybridizes to a primer. In

one embodiment, the linkers are identical to one another and hybridize to the same

primer. In a preferred embodiment, the linkers are different from each other such

- 22 -

that each hybridizes to a different primer.

   The double-stranded linker is ligated to the immobilized PCR product. The linker may incorporate the Type IIS restriction enzyme site or it may incorporate only a portion of the site. In this case, the linker will be designed such

5 that ligation of the linker to the restricted DNA will reconstitute the Type IIS restriction site. In a preferred embodiment, the linker incorporates a *Bpm*I site. Linker ligation is well-known and may be accomplished by any method known in the art. After ligation, the immobilized PCR product is isolated from the free linkers by any method known in the art. See, e.g., Velcelescu et al., Science 270: 484-487,

10 1995; Powell, Nucleic Acids Research 14: 3445-3446, 1998; Sambrook et al., pp. F.8-F.10, 1989.

   Type IIS restriction enzymes cleave at a defined distance up to 20 basepairs away from their asymmetric recognition sites. Type IIS restriction enzymes that are commercially available include enzymes that leave 5' overhangs

15 and those that leave 3' overhangs as double-stranded DNA products. Some enzymes of the former class include: *Bsm*FI (10/14), *Bst*71I (8/12), and *Fok*I (9/13), where the number in parentheses indicate the cleavage position on the same DNA strand as the recognition sequence/cleavage position on the complementary DNA strand. Enzymes of the latter class include: *Bpm*I (16/14), *Bsg*I (16/14),

20 *Eco*57I (16/14) and *Gsu*I (16/14). The 3' overhang left by these enzymes must be removed for a blunt ligation (*Step IV*). Therefore, enzymes that cleave at positions 16/14 result in a 14 base-pair tag. Other enzymes that cut at a more distal position could create a larger tag. For instance, *Mme*I (20/18) leaves a 3' overhang, but is not commercially available (Tucholski, J. et al., 1995, *Mme*I, a class-IIS restriction

25 endonuclease: purification and characterization. *Gene* 157: 87-92).

Step IV. Digestion of the Product with Type IIS Restriction Enzyme

   The product is digested with the appropriate Type IIS restriction

- 23 -

enzyme to release a DNA fragment from the anchoring bead and produce a short
hybrid DNA fragment containing a portion of the polymorphic region of the DNA
of interest (the tag) and the linker DNA. After digestion, the DNA must be either
filled in or digested to create blunt ends. If the Type IIS restriction enzyme

5   produces a 3' overhang, the fragment is digested with T4 DNA polymerase to
remove the 3' overhang. If the Type IIS restriction enzyme produces a 5' overhang,
the overhang must be filled in using the appropriate deoxynucleotides and the
Klenow fragment of DNA polymerase I. The DNA fragment is separated from the
rest of the immobilized PCR product. In a preferred embodiment, the two pools of

10   immobilized PCR products are digested with *BpmI* to release the polymorphic
markers and digested with T4 DNA polymerase to create blunt ends (Figure 2).

Step V. Ligation of Tags and PCR Amplification of Resulting Ditags

            The tags are blunt-end ligated to one another using methods well-
known in the art to form ditags. See, e.g., Sambrook et al., 1989; Velcelescu et al.,

15   Science 270: 484-487, 1995. The ditags are subsequently amplified by PCR using
primers that are unique to the linkers used in Step III. In a preferred embodiment,
the primers are different from one another if the linkers used in Step III were
different from one another. Alternatively, the primers may be the same if the linkers
used in Step III were identical to one another. The number of PCR amplification

20   reactions will vary depending upon the amount of DNA present in the starting
material. If there is a large amount of DNA, then only one PCR amplification
reaction, wherein each reaction comprises from approximately 15-30 cycles, will be
required at this step. If the starting amount of DNA is low, then more than one
PCR amplification reaction may be required at this step.

25   *Step VI. Cleavage of the Ditags and Ligation to Form Ditag Concatemers*

            The ditags are cleaved with the four-base restriction enzyme used in

- 24 -

Step II. The products are then ligated to create ditag concatemers. In one
embodiment, the ditag concatemers range from 2 to 200 ditags. In a more preferred
embodiment, the concatemer comprises 20-50 polymorphic tags. The concatemer
may be sequenced directly, or may be cloned into a sequencing vector. Using a 96-
5   channel capillary DNA sequencer, about 12,000 tags could be easily analyzed in one
day. Alternatively, the concatemers may be sequenced manually.

Methods to Analyze Marker Data

The invention is directed toward methods of analyzing the genetic
diversity of a population in a sample. Each population that is analyzed will have its
10   own unique set of different organisms or genes. The data set that is captured from
each sample should recapitulate the genetic structure in a survey format to include a
marker for each gene or organism and the relative abundance of each gene or
organism in the population as a whole. The markers for a particular population
form a marker diversity profiles (MDPs), that may be entered into a database. See,
15   e.g., Fig. 8 which shows one schematic for generating such a database. The method
by which the data are captured is not critical as long as it produces an accurate
representation of each population.

In one aspect of the method, MDPs are entered into a database. In a
preferred embodiment, the database is kept in a computer-readable form, such as on
20   a diskette, on a non-removable disk, on a network server, or on the World Wide
Web. However, the method by which the data are captured is not critical as long as
it produces an accurate representation of each microbial community.

Artificial intelligence (AI) systems can perform many data
management and analysis functions. Examples of AI systems include expert systems
25   and neural networks. Expert systems analyze data according to a knowledge base
in conjunction with a resident database. Neural networks are comprised of
interconnected processing units that can interpret multiple input signals and generate

- 25 -

a single output signal. AI systems are ideally suited for analyzing complex

biological systems including populations through the use of deduction protocols.

A marker may be correlated with a particular condition or with

another marker. See, e.g, Fig. 9 for a schematic of the steps involved in determining

5   particular parameters associated with an MDP and Fig. 10, which shows a

schematic for generating a marker diversity matrix database. A condition or state

may be an environmental condition such as pH, temperature, salinity, or the

presence or absence of an organic or inorganic compound such as hydrocarbons,

nitrates or mineral deposits. A condition may be a physiological or medical

10   condition such as an acute or chronic disease state, physiological state,

developmental stage or associated with a particular body tissue or fluid. Information

regarding all known parameters associated with the samples will also be saved

together with the MDPs.

Each MDP is composed of markers which represent a small number,

15   more preferably one, species or gene. For instance, in the case of Example 1, each

marker would be comprised of a 12 base-pair polymorphic 16S rDNA sequence.

Such parameters that relate to environmental samples may include inorganic

components (obtained through atomic adsorption analysis), organic components

(obtained through GC-MS or LC-MS), grain size analysis, pH, and salinity.

20   Parameters that relate to medical samples would include, but are not limited to, a

complete medical history of the donor. See, e.g., Fig. 11, which shows a schematic

for mapping applications using marker diversity profiles.

In another aspect of the invention, MDPs are collected for a time

course, and each time point is one of the parameters included. Time courses may be

25   useful for tracking changes over time for a wide variety of indication. For example,

time courses may be useful for tracking the progression of a disease, during

environmental remediation efforts, and during oilfield production.

In another aspect of the invention, MDPs are collected in various

- 26 -

distinct locations, such as in various geographical locations or in various tissues of the body. Comparison of MDPs compiled from various distinct locations are useful for distinguishing changes between these various locations, which may be indicative of particular environmental conditions or disease states.

5   Comparison of marker diversity profiles can reveal trends in populations either relative to time or to geographical location. In the latter case, comparisons of microbial populations can resolve spacial information about the environment that would otherwise be undetected. Examples of such information include migration patterns of water, organic compounds and minerals. For instance,

10   placer deposits of minerals are caused by the action of both water and wind causing the minerals to migrate from a lode deposit at one location only to deposited at another location. The migration of such minerals may leave a detectable trace upon the microbial populations in the path of migration. Physical attributes of the environment could also be detected such as structures, formations and fault lines. It

15   is commonly understood that faults offer a significant vertical migration route for gases such as methane which is known to be differentially utilized by microbes. By combining MDP data with geographical coordinates such as elevation, longitude and latitude that can easily be obtained with global positioning system devices, it is possible to create maps delineating the distribution of various microbes in the

20   environment.

          Correlation analyses between one marker and all the other markers in the database will reveal pairs of markers that have a propensity to coincide. This process can be repeated in an iterative manner for all markers to produce a matrix of correlation coefficients between all observed markers. Figure 5 shows a scatter plot

25   for two pairs of markers with one of the pairs exhibiting a high degree of correlation. This approach can also be used to create a dendrogram that reflects the relative level of correlation between each marker. Therefore, at any chosen level of correlation, all of the observed markers can be divided into groups where the

- 27 -

markers in each group share the same level of correlation with each other member
of the group. If a high correlation coefficient value is chosen (e.g. 0.8), the markers
of each group would, more often than not, be found in the same sample. Thus, this
exercise will divide a given population into groups of genes or organisms that have

5   a propensity to co-localize with each other. In one preferred embodiment, the
exercise will divide a microbial community into groups of microbes that have a
propensity to co-localize.

Correlation analysis between a marker (variable 1) and a sample
parameter (variable 2) will identify markers whose presence often, or invariably,

10   coincides with a component present in the samples. Some types of relationships
between markers and sample components (or parameters) are shown in Figure 6. A
strong correlation value between a marker and sample parameter would allow
predictions to be made about the abundance of either variable (marker or sample
parameter) as long as one of the variables is known.

15   In some cases, a marker will not be specific for a single species or
gene. For example, the tag sequence that would be identified by the approach
depicted in Example 1 would be identical for *Denitrobacter permanens* and
Legionella anisa. In the cases where a significant correlation is found between a
marker and sample parameter of interest, the preferred action is to use the tag

20   sequence information to identify the complete gene sequence. The sequence can
then be used to identify the species and to identify species-specific probes to verify
the correlation. One may do this using methods known in the art, such as by PCR
or by hybridization to DNA molecules isolated from the sample of interest, followed
by sequencing or other method of analysis.

25   Species-specific probes that are identified from markers with a
robust correlation to a sample parameter of interest can then be utilized as a
diagnostic, or to prospect for the parameter of interest. Such assays would
preferably be PCR-based and would be highly sensitive, rapid and inexpensive. In a

- 28 -

preferred embodiment, a marker identified by these methods may be used as a hybridization probe to identify a larger piece of the DNA from which the marker is derived. The sequence of the larger DNA molecule can then be used to design primers that will specifically hybridize to the DNA molecule of interest and which can be used to specifically amplify the DNA molecule by PCR. Alternatively, one may use a hybridization-based assay using a probe that binds specifically to the DNA molecule of interest. Using specific primers or probes are especially useful for quickly determining whether a large number of samples contains the DNA molecule that correlates to the parameter of interest.

In a preferred embodiment, a marker that correlates with a desired parameter is identified. The marker may be identified using SARD, or may be identified using another method, such as restriction fragment length polymorphisms (RFLP) or terminal restriction fragment length polymorphisms (T-RFLP; Liu et al., Applied and Environmental Microbiology 63: 4516-4522, 1997). A method such as denaturing gradient gel electrophoresis (DGGE) may be used to identify size differences. In a preferred embodiment, SARD is used to identify the marker. Other samples are screened to determine whether they have the marker of interest. In a preferred embodiment, the screen used is PCR or hybridization, more preferably PCR. In an even more preferred embodiment, the marker correlates with the presence of mineral deposits or petroleum reserves.

Correlation analysis between MDPs (MDPn, variable 1; MDPn+1, variable 2) can reveal the relative similarities between samples. Samples taken from the same individual or from proximal environmental sites that have similar composition, are expected to show a robust correlation coefficient (Figure 7A). However, samples that share only one or a few parameters in common are expected not to show a significant correlation when all of the markers are considered (Figure 7B). By incorporating the knowledge learned from the Marker-Marker and Marker-Parameter correlations, MDPs can be compared using either individual

- 29 -

markers or preferably, subsets of correlated markers in the analysis (Figure 7C).
This approach can eliminate much of the noise and enable one to identify hidden
relationships.

Correlation analyses may be performed by any method or calculation

5   known in the art. Correlation analyses for r and r^2 may be performed as described
by M. J. Schmidt in Understanding and Using Statistics, 1975 (D. C. Health
and Company), pages 131-147. The degree of correlation for r may be defined as
follows:

1.0	Perfect
0.8-0.99	High
0.5-0.7	Moderate
0.3-0.4	Low
0.1-0.2	Negligible

10

In one embodiment, the correlation between two markers or between

15   a marker and a parameter is at least low (r is 0.3-0.4). In a preferred embodiment,
the correlation is at least moderate (r is 0.5 to 0.7). In a more preferred
embodiment, the correlation is high (r is 0.8 to 0.99).

With the development of numerous genomic technologies for
analyzing complex sets of nucleic acids, we have the opportunity to begin to catalog

20   the reservoir of microbial, and hence, metabolic diversity. Since the proliferation of
a microbe in a given location will depend upon the presence of the requisite
metabolic nutrients, information as to the abundance of that microbe can serve as a
biosensor for a given set of parameters. When viewed as a whole, the microbial
community structure in a given location will hold intrinsic biosensor potential for a

25   wide range of parameters. The predictive reliability of the data from a complete
microbial community will also be significantly increased. For example, if a given
microbe were present in 50% of soil samples taken above petroleum reservoirs and
were found nowhere else, then the presence of ten such microbes would create a

- 30 -

predictive value with 99.9% accuracy.

Applications of the Invention

Geochemical and Mineral Exploration

        The methods described in this invention have several benefits over

5   existing technologies. For instance, in the area of geochemical exploration, genomic

rDNA-based assays potentially will be able to resolve an extensive set of

geochemical parameters of interest to the petroleum and mining industries.

Currently, many different technologies are required to measure these parameters.

Because this invention is based upon a universal measure, nucleic acid detection, it

10   can greatly reduce instrumentation and sample outsourcing costs.

        Oil and gas reservoirs are located well beneath the earth's surface at

depths from a few hundred feet to more than 10,000 feet. When oil is formed, it

undergoes a migration in which one of two things take place. The oil may continue

to migrate until it ultimately reaches the surface, where it evaporates over time.

15   Alternatively, its migration may be blocked by an impermeable structure, a so-called

"trap". Geophysical methods (such as three-dimensional seismic methods) for

petroleum exploration relies on finding these trap structures with the hope that they

contain oil.

        Crude oil is made up of a variety of hydrocarbon chain lengths. The

20   lightest hydrocarbons (namely methane, ethane, propane and butane) are often able

to diffuse through the trap structures and, as a result of pressure gradients, undergo

a vertical migration to the surface. Certain microbes present at the surface or in the

surface layer are able to utilize these migrating hydrocarbons, which occasionally

results in mineralogical changes that are detectable at the surface. Thus, these

25   migrating hydrocarbons would be expected to affect microbial populations, such

that the ability to determine the genetic diversity of a microbial population may

reveal microbial signatures that are indicative of the presence of oil.

- 31 -

Recent advances in microfluidics in the genomics industry have resulted in the development of instruments that can detect specific nucleic acids within a few minutes. Utilizing such instruments will enable measurements to be made in the field for a variety of parameters. In contrast, conventional chemical assays require laboratory analysis and interpretation.

Biosensors have been created that are able to detect hydrocarbons present at, or below, the level of detection of sophisticated GC-MS analytical instrumentation (Sticher, P. et al., 1997, Development and characterization of a whole-cell bioluminescent sensor for bioavailable middle-chain alkanes in contaminated groundwater samples. *Appl Environ Microbiol.* **63**:4053-4060). Sticher et al. demonstrated that by using single reporter gene in a genetically engineered microbe comprising a reporter gene was able to sense extremely small changes in their environment in response to an acute treatment with a particular hydrocarbon. The instant invention could document the effect upon a population comprising thousands of microbes over geologic time and thus, has the potential of being more sensitive than current analytical instruments.

This invention may also be used to create a survey of biological entities that is limited only by the prerequisite that these entities contain nucleic acids that are arranged in regions that are conserved and regions that are polymorphic when compared to sequences from related organisms. Some additional examples of the application of this invention are described below.

Oil and Gas Reservoir Development

In addition to the application of this invention in petroleum exploration, this invention could also be useful in the development of oil and gas reservoirs. Several properties of oil reservoirs that directly affect the commercial viability of the reservoir are modulated at some level by microbes. Hydrogen sulfide is sometimes present in crude oil and can render otherwise 'sweet' oil into 'sour'

- 32 -

oil. In addition to its corrosive effect on oilfield equipment, H₂S also poses risk to the workers and significantly reduces the value of an oil reservoir since a washing plant must be installed to remove the gas. The levels of H₂S can change during the development of a reservoir and is now thought to be the result of sulfate-reducing
5   bacteria (Leu, J.-Y. et al., 1999, The same species of sulphate-reducing *Desulfomicrobium* occur in different oil field environments in the north sea. *Lett. Appl. Microbiol.* 29(4):246-252). By identifying the presence of microbes that could lead to H₂S production, the valuation of new reservoirs and the resulting developmental strategies could be made more effective.
10   Crude oil and natural gas are composed of a complex mixture of hydrocarbons including straight chain hydrocarbons of lengths generally between 2-40 carbon atoms. The shorter chain-length hydrocarbons are more valuable (e.g. gasoline, C_4-C_{10}). In some oil reservoirs, the lighter hydrocarbons are selectively removed either during or prior to development of the reservoir. Microbes have
15   been suspected to play a role in this process since the shorter chain-length hydrocarbons are more bioavailable. This invention could identify microbes that are involved in this process and therefore make predictions as to the susceptibility of certain reservoirs to the depletion of short chain hydrocarbons. This invention may also be able to identify microbes capable of shortening long chain hydrocarbons
20   thereby increasing the value of existing reservoirs.

Insect and Parasite Detection

             The significant negative impact insects can have on agriculture is widely known. Insects can also serve as vectors for the transmission of many disease causing microbes. Numerous microbe-insect relationships have been
25   described. For example, the bacterial genus *Wolbachia* is found associated with many species of ants and has been shown to alter sex determination and fecundity in the host (Wenseleers, T. et al., 1998, Widespread occurrence of the micro-organism

- 33 -

Wolbachia in ants. *Proc. R. Soc. Lond. B. Biol. Sci.* 265(1404):1447-52). In
addition, many intracellular endosymbiotic bacterial species have been identified in
ants (Schroder, D. et al., 1996, Intracellular endosymbiotic bacteria of *Camponotus*
species (carpenter ants): systematics, evolution and ultrastructural characterization.
5   *Mol. Microbiol.* 21:479-89). Most, if not all, insects probably have species-specific
intimate relationships with microbes which could represent could represent an
Achilles heel for the control of insect populations. The invention described in this
application could provide a means to identify microbes that modulate the well-being
of a given insect species.

10   The identification of microbes that are specifically associated with a
given insect could also potentially serve as a the basis for a highly sensitive test for
the presence of the insect. For instance, current methods to identify the presence of
termites in wooden structures is based on visual inspection and is largely
inadequate. A test for the presence of a termite-associated microbe that is based on
15   PCR-amplification would be both non-invasive and highly sensitive.

            Further, the ability to create comprehensive inventories of microbial
diversity has several applications that relate to microbial ecology that would have
utility in the agricultural industry. For instance, the agriculture industry utilizes
enormous amounts of pesticides prophylactically to prevent loss of crops. This
20   invention provides the ability to perform comprehensive surveys of microbial
populations and could lead to predictions as to the susceptibility of a given field to
particular plant pathogens. This knowledge could lead to a better strategy of
pesticide applications.

            Further, surveying microbial diversity in an environment such as a
25   agricultural field at various times of the year would reveal the dynamic changes in
microbial populations that occur as a result of seasonal fluctuations (temperature
and moisture), pesticide application and the proliferation of certain organisms.

            Comparison of the diversity profiles taken from the same or similar

- 34 -

site at different times would reveal interactions between species in a population. These productive interactions may manifest themselves either in the increase or decrease in the representation of one marker relative to the decrease or increase in the representation of a second marker, respectively.

5   Such information could provide an early warning of the proliferation of a pest species as well as the identification of species that are pathogenic to a pest species. These pathogenic organisms may have value either as a biological agent to control proliferation of pest species and/or as a source for genes or compounds that would act as pesticides. This phenomenon is not be limited to microbe-microbe

10   interactions. The eggs as well as larval stages of insects likely interact with soil microbes and create a detectable impact upon microbial populations. An example of such an organism is *Bacillus thuringensis*, which itself is commonly used in organic farming as an insecticidal agent. The gene responsible for this insecticidal activity (Bt toxin) has been widely used to create transgenic plants with resistance to insect

15   attack.

Bioremediation

A considerable amount of effort has gone into the development of methods for microbe-based removal of chemicals from the environment. Such chemicals include heavy metals, polynuclear aromatics (PNAs), halogenated

20   aromatics, crude oil, and a variety of other organic compounds such as MTBE. Regulatory considerations for the release of micro-organisms into the environment have re-directed efforts towards identifying and augmenting the growth of endemic organisms that have the capability to metabolize or remove compounds of interest from the environment.

25   The present invention can facilitate bioremediation efforts in two ways. First, organisms can be identified at a given site that have either been previously shown to be capable of removing compounds of interest, or that have

- 35 -

significant likelihood of having the capacity metabolize the relevant compounds
based upon its coincidence in the environment with the chemical in numerous
geological settings. Secondly, this invention can identify trends of soil types and
particular microbial species. In this case, the correlations that are drawn from the

5   database between microbial distribution and soil types can be coupled with the
existing knowledge base of geochemistry. For instance, the USGS provides many
publically available maps describing numerous geochemical and geophysical
parameters. Extrapolations of the distribution of microbial species can be made to
the regional, and possibly worldwide, level. These extrapolated microbial

10   distributions could serve as the basis for site-specific treatment regimens to augment
the growth of certain relevant species without first performing a microbial survey.

Immunological Applications

Methods could be applied to map immunoglobulin and TCR gene
rearrangements, which occur normally during B and T cell differentiation. These

15   rearrangements could provide a profile of an individual's immunoglobulin and TCR
diversity that could be correlated with medical history. This type of analysis might
lead to the early diagnosis of certain conditions or diseases and allow for a more
proactive and early treatment. Further, samples of immune cells could be isolated
from individuals or certain body fluids or tissues to identify potential

20   immunoglobulin or TCR profiles that may be correlated with particular diseases,
particularly autoimmune diseases. For instance, it has been demonstrated that T
cells expressing particular TCR subtypes are found in higher levels in the synovial
fluid of individuals suffering from rheumatoid arthritis. The methods of the instant
invention may be used to identify other correlative immunoglobulin or TCR profiles

25   in autoimmune and other diseases.

- 36 -

Virus Detection

The human genome contains thousands of copies of human endogenous retroviruses (HERVs) that make up as many as 1% of the human genome (Sverdlov, E.D. 1998, Perpetually mobile footprints of ancient infections in

5   human genome. *FEBS Lett.* 428:1-6). These sequences are thought to be remnants of infections that occurred millions of years ago. These sequences can transpose to other locations in the human genome and may be responsible for disease-susceptibility in certain human populations (Dawkins, R. et al., 1999, Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and

10   disease. *Immunol. Rev.* 167:275-304). This invention could be used to survey HERV polymorphic sequences and determine whether they correlate with a variety of clinical parameters.

Forensic Science

This invention, particularly SARD analysis, can be used in forensic

15   applications as well. Studies over the past ten to twenty years have focused upon the changes that occur in a person's or animal's body after death. Many of these changes involve changes in microbial populations that occur during decomposition of the body. Changes in microbial populations have been correlated with length of time that a person has been dead and the conditions that the body experienced after

20   death, e.g., heat, sun exposure, partial or complete burial, rain, etc. SARD analyses would permit forensic scientists to quickly and accurately determine the size and type of microbial populations, which in turn may be used to determine more accurate times of death as well as conditions that the body may have been exposed to.

25   *Other Applications*

This approach can be utilized for any polymorphic region in a

- 37 -

genome, whether microbial, viral or eukaryotic that is flanked by conserved DNA
sequences. This method also need not be restricted to genes. The DNA sequence
of intergenic regions of genomes are not under as high a level of selective pressure
and thus, represent highly polymorphic DNA sequence. One example of such a

5   region is the intergenic region between the large (23S) and small (16S) rDNA
subunits coding regions. The above-described methods may be used to distinguish
members of a population based upon size differences of the intergenic region
between the 16S-23S or 23S-5S rDNA genes. The spacer region between these
genes has been found to be hypervariable in microbial populations. In the case of

10   the 16S-23S intergenic region, the spacer size ranges between about 200-1500
base-pairs depending the presence or absence of various tRNA genes. (Nour M.
1998, 16S-23S and 23S-5S intergenic spacer regions of lactobacilli: nucleotide
sequence, secondary structure and comparative analysis, *Res. Microbiol.*
149(6):433-448; Berthier F. et al., 1998, Rapid species identification within two

15   groups of closely related lactobacilli using PCR primers that target the 16S/23S
rRNA spacer region, *FEMS Microbiol Lett.* **161**(1):97-106; Tilsala-Timisjarvi A. et
al., 1997, Development of oligonucleotide primers from the 16S-23S rRNA
intergenic sequences for identifying different dairy and probiotic lactic acid bacteria
by PCR *Int. J. Food Microbiol.* **35**(1):49-56). Both of these rDNA genes are

20   transcribed on the same operon. Therefore, the conserved regions of the rDNA
coding sequence of these subunits can be utilized to amplify the intergenic regions.

In order that this invention may be better understood, the following
examples are set forth. These examples are for purposes of illustration only and are
not to be construed as limiting the scope of the invention in any manner.

- 38 -

EXAMPLE 1

Serial Analysis of rDNA Polymorphic Tags from the Domain *Eubacteria*

A sample comprising environmental bacteria was obtained and total
DNA was extracted from the sample. To amplify the DNA, PCR was performed by
mixing 5 μL 10X Advantage2 reaction buffer (Clontech), 2.5 μL dNTPs, 5 μL 8
μM TX9/TX16 primers, 50 ng of sample DNA, 0.5 μL Advantage2 Taq polymerase
(Clontech) and water to 50 μL. Primer TX9 was biotinylated. The primers are
shown in Figure 12 and a general strategy of this example is shown in Figure 13.
The reaction mixture was then subjected to PCR under the following conditions:

          a) 94°C for 5 minutes;
          b) 94°C for 1 minute, 10 seconds;
          c) 55°C for 50 seconds;
          d) 68°C for 1 minute;
          e) repeat steps (b) through (d) 20 times; and
          f) 68°C for 2 minutes.

The approximately 600 basepair PCR product was gel purified using
1% agarose and a Qiaquick kit (Qiagen) according to manufacturer's instructions.
The DNA was eluted with 50 μL Tris-EDTA (TE; 10 mM Tris pH 8; 1 mM
EDTA). To restrict the amplified DNA, 10X New England Biolabs Buffer #2
(NEB#2) was added to bring the concentration of the buffer to 1X and 25 units of
*Alu*I was added. The reaction mixture was incubated for 2 hours at 37°C; 25
additional units of *Alu*I was added and the reaction mixture was incubated for a
further one hour followed by inactivation of the enzyme at 65°C for 20 minutes.

In order to immobilize the restricted DNA fragment to a bead, an
equal volume of 2X BW buffer (2X BW = 10 mM Tris pH 7.5; 1 mM EDTA; 2 M
NaCl) was added to the reaction mixture, followed by addition of 50 μL washed M-
280 Streptavidin (SA) beads (Dynal). The reaction mixture was incubated for 20
minutes at room temperature. The beads were then washed twice in 1X BW and

- 39 -

twice in wash buffer (10 mM Tris pH 8; 10 mM MgSO₄; 50 mM NaCl). During the last wash step, the beads were split into two pools.

To add the linkers to the immobilized DNA, one pool of beads were resuspended in 4 µL of 10 µM linker TX-12/13 (a double-stranded DNA molecule
5   comprising primers TX-012 and TX-013) to form pool A, while the other pool (pool B) were resuspended in 4 µL of 10 µM linker TX-14/15 (a double-stranded DNA molecule comprising primers TX-014 and TX-015) was added to the other pool. 36 µL of T4 ligase mix (4 µL 10X ligase buffer; 32 µL water; 0.2 µL T4 ligase [400 U]) were added to the linker/bead mixture and the mixture was
10 incubated overnight at 16°C.

The DNA molecule attached to the streptavidin beads was then incubated with BpmI, which recognizes its specific restriction sequence (GAGGTC) in the DNA molecule. BpmI cleaves the DNA such that it releases the streptavidin bead from the DNA molecule and incorporates a part of the polymorphic region of
15 the DNA molecule first amplified. The beads were washed twice with 0.5 mL each 1X BW and twice with wash buffer. The beads were resuspended in 10 µL BpmI mix (1 µL 10X NEB#3 [New England Biolabs]; 1 µL 1 µg/µL bovine serum albumin; 8 µL water and 1 µL BpmI [New England Biolabs]). The reaction mixture was incubated at 37°C for 2 hours and then at 65°C to inactivate the enzyme. The
20 supernatant containing the DNA tags was then isolated.

In order to remove the 3' overhang on the DNA tags and make them blunt ended, to the supernatant (10 µL) on ice was added 10 µL T4 polymerase mix (1 µL 10X NEB#2; 0.5 µL 4 mM dNTPs; 8.5 µL water and 0.33 µL T4 polymerase [1 U; New England Biolabs]). The reaction was incubated at 12°C for 20 minutes
25 and at 65°C for 20 minutes to inactivate the polymerase. In order to form ditags, pool A and pool B were recombined to give a total volume of 40 µL. Four µL 10 mM rATP and 0.2 µL T4 DNA ligase (400 U) was added and the reaction was incubated 4 hours to overnight at 16°C.

- 40 -

As an intermediate amplification step, the ditags were then amplified in a 300 μL PCR reaction (30 μL 10X Advantage2 Taq buffer; 15 μL 4 mM dNTPs; 30 μL 8 μM TX111/TX121 primer mix; 3 μL Advantage2 Taq polymerase; 3 μL ditag template; 219 μL water). The 300 μL reaction mix was split into three 100 μL reactions and amplified using the following conditions:

5

      a) 94°C for 5 minutes;

      b) 94°C for 30 seconds;

      c) 56°C for 30 seconds;

      d) 68°C for 40 seconds;

10

      e) repeat steps (b) through (d) for 15 cycles; and

      f) 68° for 2 minutes.

After amplification, three volumes (900 μL) QG buffer (Qiagen) and one volume (300 μL) isopropanol was added, the mixture was bound to a Gel Extraction Spin Column (Qiagen) and the DNA was eluted with 50 μL TE.

15

The amplified DNA was then subjected to non-denaturing polyacrylamide gel electrophoresis (PAGE; Novex 1 mm 10% TBS-PAGE gel). The gel was stained with 5 μg/mL ethidium bromide, the approximately 106 basepair ditags were excised from the gel, the excised gel was fragmented and 0.3 mL of TE was added to the gel and incubated at 65°C for 30 minutes. The gel/TE

20

mixture was transferred to a miniprep spin column (Qiagen) and the eluate containing the amplified ditags was collected (the ditag template). This amplified DNA was called the P300 ditag template.

In order to determine the optimal number of PCR cycles for large-scale amplification (PCR cycle titration), 6.4 μL P300 ditag template was mixed

25

with 8 μL 10X Advantage2 Taq buffer, 4 μL 4 mM dNTPS, 8 μL TX111/TX121 primers, 0.8 μL Advantage2 Taq polymerase and 52.8 μL water. The reaction mixture was split into three 25 μL reactions and amplified for 6, 8 or 10 cycles using the PCR conditions described above for the intermediate amplification step. After

- 41 -

amplification, the DNA products were desalted and purified using a Qiagen Gel
Extraction Spin Column as described above except that only 75 µL QG buffer and
25 µL isopropanol were used, and the DNA products were eluted with 20 µL TE.
The cycle number that produced the largest 106 basepair ditag yield without any

5   detectable vertical smearing by 10% TBS-PAGE analysis was used for large-scale
amplification.

        For large-scale ditag amplification, 180 µL 10X Advantage2 Taq
buffer, 90 µL 4mM dNTPs, 180 µL 8 µM TX111/TX121 mixture, 18 µL
Advantage2 Taq polymerase, 144 µL P300 ditag template and 1188 µL water was

10   mixed together and then split into 18 100 µL reactions. PCR was performed as
described immediately above using the number of cycles that were determined from
the previous titration step. The PCR reactions were desalted by adding three
volumes QG buffer and one volume isopropanol to the reactions and passed through
a total of four Qiagen Gel Extraction Spin Columns. Each column was eluted with

15   50 µL TE. The samples were subjected to non-denaturing preparative PAGE
(Novex 1.5 mm 10% TBS-PAGE). The DNA bands were stained and excised as
above and split into three tubes. The polyacrylamide was fragmented and eluted
using TE as described above and the DNA was purified by passing the contents of
the gel/TE mixture through a Qiagen Plasmid Spin Miniprep column as described

20   above.

        To precipitate the DNA, 1 µL glycogen, 150 µL 10M ammonium
acetate and 1125 µL ethanol was added to each eluate of approximately 300 µL.
The mixture was incubated at -80°C for 20 minutes and microfuged at 13,000 rpm
at 4°C for 15 minutes. The pellet was washed with 1 mL 70% ethanol and dried at

25   37°C for 5 minutes.

        The DNA pellets were resuspended in 150 µL AluI mix (15 µL
NEB#2, 135 µL water and 15 µL AluI [150 U]). The reaction mixture was
incubated at 37°C for 2 hours. 150 µL of 2X BW buffer was added and the sample

- 42 -

was transferred to a new tube containing 150 packed SA beads. The beads were
incubated with the DNA mixture for 30 minutes at room temperature and the beads
were pelleted magnetically. The supernatant was removed, extracted with 250 μL
phenol/chloroform/isoamyl alcohol, and the aqueous phase was precipitated with

5   150 μL 10M ammonium acetate and 1125 μL ethanol overnight at -80°C. The
DNA was centrifuged at 13,000 rpm at 4°C for 15 minutes, the pellet washed with
70% ethanol and dried. This step removes any free linkers that were present as well
as any incompletely digested DNA products that were still bound to the biotinylated
linkers.

10   In order to concatenate the ditags, the DNA pellet was resuspended
in 10 μL ligase mix (1 μL 10X ligase buffer; 9 μL water and 1 μL T4 ligase [100
U]) and incubated at 16°C for 30 minutes. The DNA was precipitated by added 40
μL TE, 25 μL 10M ammonium acetate and 180 μL ethanol. The DNA was
precipitated for 15 minutes at -80°C and centrifuged, washed and dried as described

15   above.

             The concatenated ditags were purified by resuspending the pellet in
1X TBE loading buffer and subjecting the sample to non-denaturing PAGE (8%
TBE-PAGE). The area between 0.5 to 1.2 kb was excised and the DNA was eluted
from the gel into 300 μL TE as described above. The DNA was precipitated using

20   1 μL glycogen, 150 μL 10M ammonium acetate and 1125 μL ethanol. The DNA
was precipitated, centrifuged, washed and dried as before.

             The concatenated ditags were cloned in pUC19 by resuspending the
DNA pellet in 3 μL SmaI-cut pUC 19 (approximately 100 ng plasmid DNA) and
adding 7 μL T4 ligase mix (1 μL 10X ligase buffer, 6 μL water and 0.2 μL T4 ligase

25   [400 U]). The plasmid/ditag mixture was incubated for 2 hours at 16°C and 1 μL of
the mixture was used to transform 100 μL chemically-competent DH10B. Amp-
resistant transformants were screened by PCR using pUC/m13 forward/reverse
17mers as PCR primers. Transformants containing concatemerized ditags were

- 43 -

then sequenced.

EXAMPLE 2

SARD Analysis of a Defined Population

SARD was performed essentially as described in Example 1. In this

5   example, commercially available bacterial genomic DNA samples were mixed at an

equal concentration (weight/volume). The bacterial DNA samples used included

Bacillus subtilis, Clostridium perfringens, Escherchia coli, Lactococcus lactis and

Streptomyces coelicolor. Equal volumes of the DNA samples (total genomic DNA)

were mixed at a concentration of 50 ng/μL each. Chart I shows the size of the

10   genomes of each bacterial species, the number of 16S rDNA copies per genome and

the molar percentage of 16S copies for each bacterial species in the total DNA

sample.

Chart I

Bacteria	Genome (Mb)	16S Copies/Genome	Molar %
B. subtilis	4.2	10	17.2
C. perfringens	4.4	10	16.4
E. coli	4.6	7	11.0
L. lactis	2.4	10	30.7
S. coelicolor	8.0	6	5.4

20   After SARD, 120 tags were sequenced. Chart II shows the expected

number and percentage of tags compared to the observed number and percentage of

tags for the population. See also Fig. 14.

- 44 -

Chart II

Bacteria	Expected Tag Number	Observed Tag Number	Expected Tag Percentage	Observed Tag Percentage
B. subtilis	21	44	17.2%	36.7%
C. perfringens	20	18	16.4%	15.0%
E. coli	13	8	11.0%	6.7%
L. lactis	37	35	30.7%	29.2%
S. coelicolor	6	6	5.4%	5.0%

Each of the rDNA genes from these species produced a SARD tag
that was distinguishable from the other members of the set. As can be observed,
approximately twice as many tags that corresponded to 16S rDNA from *B. subtilis*
were found than was expected based upon the molar percentage of 16S rDNA from
B. subtilis. The observation that *B. subtilis* appeared to be twice as abundant as
was expected has been reported previously. Farrelly et al. (Farrelly, V. F. et al.,
1995, Effect of genome size and rrn gene copy number on PCR amplification of 16S
rRNA genes from a mixture of bacterial species. Appl. Environ. Microbiol.
61(7):2798-2801) described similar results when PCR amplifying the 16S rDNA
gene from mixed populations of genomic DNA. They concluded that this
phenomenon was the result of the tandem organization fo the *rrn* operons in the *B.
subtilis* genome where multiple rDNA genes may be amplified as a single product.
The remaining tags were found at abundances close (<40% deviation) to their
expected values.

EXAMPLE 3

SARD Analysis of Environmental Bacterial Diversity

In order to demonstrate that the SARD method could be used to
survey environmental bacterial diversity, total DNA was extracted from two soil

- 45 -

samples (Wy-1 and Wy-2) taken from the Rocky Mountain Oilfield Testing Center (RMOTC, Casper, Wyoming) in October, 2000. The samples were collected about 0.5 miles apart and from a depth of 14-18 inches. The environmental DNA samples were subjected to SARD analysis as described in Example 1. In a preliminary

5   analysis, 148 tags were identified from Wy-1 and 234 tags were identified from Wy-2 (Figures 15 and 16, respectively).

        In the Wy-1 sample, 58 distinct tags were identified and the abundance of each tag varied. The most abundant tag (ATGGCTGTCGTCAGCT) made up about 34% of the population. This tag sequence is identical to many

10   bacterial sequences in GenBank and its position within the 16S rDNA gene indicates that it is located in a conserved region located distal to the targeted *Alu*I restriction site. In other words, the contributing 16S gene(s) for this tag did not contain the conserved *Alu*I site. Since the SARD tag position is dictated by the first *Alu*I site distal to the biotinylated primer used in the initial PCR reaction, it is likely that the

15   first *Alu*I site in the contributing 16S gene(s) was located downstream within a conserved region. In order to decrease the number of tags that do not contain the conserved *Alu*I site next to the polymorphic region, one may gel purify the approximately 100 basepair PCR products after the first *Alu*I restriction step. However, this may result in losing some information. Nevertheless, 39% of the tags

20   (58/148) in this set were different from each other. See Figures 15 and 17.

        The Wy-2 sample was found to contain 79 different tags out of a total of 234 tags that were examined. Thus, 34% of the tags (79/234) in this set were different from each other. See Figures 16 and 17. As in the case with Wy-1, the tag ATGGCTGTCGTCAGCT, which represents a conserved sequence in a 16S

25   rDNA gene, was most abundant and made up about 30% of the population.

        Combining the tags from the two sets reveals a total of 105 different tags from a total of 382 tags. Thus, 26 of 58 different tags (45%) in Wy-1 were not present in Wy-2. Likewise, 47 of 79 different tags (59%) in Wy-2 were not present

- 46 -

in Wy-1. The tags that were only found in one of the samples are candidates for

bacteria with indicator value for various parameters associated with each sample.

However, there was no attempt in this preliminary analysis to obtain all of the tags

present in the two samples, so it cannot be concluded that some or most of the tags

5   found in one sample are not present in the other sample. Thus, one cannot conclude

that there are tags in these two samples that are indicators for various parameters

associated with each sample. Nonetheless, a full-fledged analysis of these samples

may provide such indicators.

EXAMPLE 4

10   **Serial Analysis of rDNA Polymorphic Tags From the Domain *Archaea***

The method described in Example 1 can also be applied to the

domain *Archaea*. The domain *Archaea* is made up of two known kingdoms,

Euryarchaeota and Crenarchaeota (Pace, N. R. 1997, A molecular view of microbial

diversity and the biosphere. *Science* 276:734-740). One set of oligonucleotides and

15   restriction enzymes can be used to survey both of these domains.

In this example, the following oligonucleotides are designed: 5'

biotin-TA(CT)T(CT)CCCA(GA)GCGG(CT)(GCT)(GC)(GA)CTT(AGCT)-3'

corresponding to position 817-838 of the *Methanococcus jannaschii* 16S rDNA

gene (GenBank Accession number M59126), and (5'-GGTG(TGC)CA(GC)C(CA)

20   GCCGCGGTAA(TC)ACC(AGCT)-3' corresponding to position 457-481 of the

Methanococcus jannaschii 16S rDNA gene.

In this example, a *Bfa*I site (CTAG) is utilized that corresponds to

position 768-771 of the *Methanococcus jannaschii* 16S rDNA gene. This site is

immediately flanking a polymorphic region.

25   The SARD method was tested *in silico* using 17 representatives

from the domain Archaea (Figure 4). 15 of the SARD tags that would be identified

were all unique to this set (Table II). Two species of the genus *Methanobacterium*

- 47 -

did not possess any *Bfa*I sites in the region that would be amplified and therefore, would not produce any tags.

EXAMPLE 5

Surveying PCR-Amplified Polymorphic rDNA Regions

5   Oligonucleotide primers that are complementary to conserved regions immediately flanking a polymorphic region could be used to amplify the polymorphic DNA sequences. The oligonucleotide primer sequences could include existing or introduced restriction sites to enable subsequence cloning into a bacterial vector. By utilizing or introducing different restriction sites in each primer, the

10   restriction digested PCR products could be concatemerized in a unidirectional fashion prior to cloning. This step would allow a serial sequence analysis of multiple polymorphic regions from a single recombinant product.

EXAMPLE 6

Surveying Translation Products of 16S rDNA Polymorphic Regions

15   The polymorphic regions of 16S rDNA could also be surveyed by the parallel identification of the translation products of a given polymorphic region. Oligonucleotides could be designed such that they can serve to amplify a polymorphic region. The 5' primer would also include a T7 polymerase binding site or other polymerase binding site, a Kozak consensus sequence, an initiator ATG

20   codon and an epitope to facilitate purification. Examples of epitopes include hemagglutinin (HA), *myc*, Flag or polyhistidine. Following amplification, the products are subjected to *in vitro* transcription/translation to produce the peptide products. These peptides are purified from the cell extract and analyzed by mass spectrometry. This type of an approach has been applied to the identification of

25   mutations in the BRCA1 gene (Garvin, A. M. et al., 2000 MALDI-TOF based mutation detection using tagged in vitro synthesized peptides. *Nature Biotechnol.*

- 48 -

18:95-97).

Although this process would not readily provide DNA sequence
information from which to deduce taxonomy, it would allow for the creation of
microbial diversity profiles comprised of 'mass tags'. These mass tags could be
5   used to identify correlations between specific tags and various sample parameters.
To test the information content of this approach, a polymorphic region of the 16S
rDNA genes from the species in Figure 3 was translated *in silico*. Of the 34
polymorphic regions examined, 32 produced a tag with a unique mass in this set
(Table III).

10   Another approach to translating the amplified polymorphic regions
would be to clone and express the sequences in whole cells. For instance,
oligonucleotides could be designed to amplify polymorphic regions that include
sequences at the 5' ends that would allow for cloning by homologous recombination
in yeast. These sequences could be cloned into an expression cassette to create a
15   fusion between a secreted protein, such as alpha factor or invertase, and an epitope
to facilitate purification and the translated amino acid sequence of the rDNA
polymorphic region. Homologous recombination in yeast is quite robust and could
easily enable the isolation of 10^3-10^4 independent recombinants on a single
transformation plate. The secreted products could be isolated from the medium and
20   identified by mass spectrometry.

EXAMPLE 7

Hybridization of Microbial rDNA to Immobilized Oligonucleotides

A complication with using short oligonucleotide probes in DNA
microarrays is the instability of short oligonucleotides duplexes. A possible solution
25   to this problem is to synthesize probes that include degenerate sequences to
accommodate unknown sequences. The length of the oligonucleotide is dictated by
the number of degeneracies, or sequence permutations, that are to be

- 49 -

accommodated. For instance, a fully degenerate 9mer oligonucleotide requires 4^9 or
262,144 different oligonucleotide sequences. The efficient hybridization of 9mer
oligonucleotides is not possible using standard conditions. One solution has been to
incorporate a ligation step in the hybridization of a target sequence to a 9mer
5   oligonucleotide probe (Gunderson, K. L. et al., 1998, Mutation detection by
ligation to complete n-mer DNA arrays. *Genome Res.* 8:1142-1153).

        Another solution could be to effectively increase the length of the
oligonucleotide by including nucleotides in the primer that are not degenerate. This
approach could be applied to a survey of a microbial community by constructing
10   oligonucleotide probes that are composed of a constant region that corresponds to a
well conserved region of a 16S rDNA gene together with a degenerate sequence
that corresponds to the flanking polymorphic region.

        An example of such a collection of degenerate oligonucleotides for
the domain Bacteria could include permutations of the following primer: 5'-
15   AACGAGCGCAACCNNNNNNNNNN-3', where N indicates any nucleotide at that
position. This sequence corresponds to position 1101-1122 of the *E. coli* 16S
rDNA gene (GenBank Accession number E05133). Alternatively, the primers could
be designed such that they are composed of a mixture of constant sequence, semi-
degenerate positions (e.g. A or G) and degenerate positions (e.g. A, G, C or T).

20   All publications and patent applications cited in this specification are
herein incorporated by reference as if each individual publication or patent
application were specifically and individually indicated to be incorporated by
reference. Although the foregoing invention has been described in some detail by
way of illustration and example for purposes of clarity of understanding, it will be
25   readily apparent to those of ordinary skill in the art in light of the teachings of this
invention that certain changes and modifications may be made thereto without
departing from the spirit or scope of the appended claims.

Table I*

Species	GenBank Acc#	Tag Sequence	Position
Desulfurobacterium thermolithotrophum	AJ001049	GTCAGTTGCCGAAGCT	814-829
Uncultured Aquificales OPS132	AF027104	GTCCGTGCCGTAAGCT	810-825
Bacteroides caccae	X83951	ATGCTGTGTCGAAGCT	1021-1036
Actinomyces bovis	X81061	TTTCCGCGCCGTAGCT	834-849
Actinomyces meyeri	X82451	TTTCTGCGCCGTAGCT	828-843
Denitrobacterium detoxificans	AF079507	CCTCCGCGCCGCAGCT	788-803
Uncultured GNS bacteria BPC110	AF154084	CCCGGTAGTCCTAGCT	765-780
Uncultured GNS bacteria GCA004	AF154104	CATCGGTGCCGCAGCT	824-839
Uncultured GNS bacteria GCA112	AF154100	CGGCGGTGCCGTAGCT	826-841
Acetobacter aceti	AF127399	ACTCAGTGTCGTAGCT	782-797
Gluconobacter asaii	AB024492	ACTCAGTGTCGAAGCT	783-798
Burkholderia sp. JB1	X92188	CCTTAGTAACGAAGCT	837-852
Denitrobacter permanens	Y12639	TCGCTGTGTCGGAGCT	789-804
Desulfobacter curvatus	M34413	CTGCTGTGCCNAAGCT	861-876
Desulfobulbus sp. BG25	U85473	CCTCTGTGTCGCAGCT	854-869
Legionella anisa	X73394	ACGATGTGCACTAGCT	790-805
Benzene mineralizing clone SB-1	AF029039	TTGCTTCTCCGGAGCT	1029-1044
Escherichia coli	E05133	CGTGGCTTCCGGAGCT	848-863
Uncultured Acidobacterium Sub.Div-1	X68464	CCGCCGTGCCGAAGCT	813-828
Uncultured Acidobacterium Sub.Div-1	Z73363	CGGCTGTGCCGAAGCT	521-536
Uncultured Acidobacterium Sub.Div-1	Z73365	CCACTGTGCCGTAGCT	521-536
Uncultured Acidobacterium Sub.Div-1	Z73368	CTGCTGTGCCGCAGCT	521-536
Uncultured Acidobacterium Sub.Div-1	Z73364	CTGCCGTGCCGGAGCT	521-536
Uncultured Acidobacterium Sub.Div-1	U68659	CCAATGTGCCGGAGCT	319-334
Uncultured Acidobacterium Sub.Div-1	D26171	CCGTCGTGCCGTAGCT	779-794
Uncultured Acidobacterium Sub.Div-1	X97101	CCGTCGTGTCGTAGCT	687-702
Uncultured Acidobacterium Sub.Div-1	X97098	CTGCCGTGTCGAAGCT	798-813
Uncultured Acidobacterium Sub.Div-1	AF047646	CTCCCGTGTCGAAGCT	779-794
Uncultured Acidobacterium Sub.Div-1	AF050548	CCGCCGTGCCGGAGCT	316-331
Uncultured Acidobacterium Sub.Div-2	U68612	CTGAGGAACGAAAGCT	226-241
Uncultured Acidobacterium Sub.Div-2	Y07646	GTGTCGTCCCGGAGCT	830-845

Uncultured Acidobacterium Sub.Div-3	X97097	GGGCTGTGCCGAAGCT	804-819
Uncultured Acidobacterium Sub.Div-3	X68466	GGTCGGTGCCGGGAGCT	796-811
Uncultured Acidobacterium Sub.Div-3	X68468	GGTCGGTGCCAGAGCT	796-811
Uncultured Acidobacterium Sub.Div-3	U68648	GGTTCGTGCCGGAGCT	317-332
Uncultured Acidobacterium Sub.Div-3	X68467	TGTCTGTGCCGGAGCT	796-811
Uncultured Acidobacterium Sub.Div-3	AF013515	TATCCGTGCCGGAGCT	799-814
Uncultured Acidobacterium Sub.Div-3	AF027004	GGTCCGTGCCGGAGCT	778-793

* Sequences shown in bold with shadow indicates they are not unique to this set.

Table II

Species	GenBank Acc#	Tag Sequence	Position
Crenarchaeota			
Aeropyrum pernix	D83259	CTAGGGGGCGGGAG	614-627
Desulfurococcus mobilis	M36474	CTAGGTGTTGGGTG	856-869
Staphylothermus marinus	X99560	CTAGGTGTTGGGCG	770-783
Metallosphaera sedula	X90481	CTAGGTGTCGCGTA	756-769
Sulfolobus acidocaldarius	D14053	CTAGGTGTCGAGTA	785-798
Sulfolobus metallicus	D85519	CTAGGTGTCACGTG	744-757
Caldivirga maquilingensis	AB013926	CTAGCTGTTGGGTG	773-786
Pyrobaculum islandicum	L07511	CTAGCTGTCGGCCG	781-794
Euryarchaeota			
Archaeoglobus fulgidus	X05567	CTAGGTGTCACCGA	780-793
Archaeoglobus veneficus	Y10011	CTAGGTGTCACCGG	758-771
Haloarcula japonica	D28872	CTAGGTGTGGCGTA	762-775
Halococcus morrhuae	D11106	CTAGGTGTGGCGTT	765-778
Methanococcus jannaschii	M59126	CTAGGTGTCGCGTC	768-781
Methanobacterium bryantii	AF028688	None	
Methanobacterium subterraneum	X99045	None	
Pyrococcus abyssi	Z70246	CTAGGTGTCGGGCG	767-780
Picrophilus oshimae	X84901	CTAGCTGTAAACTC	742-755

Table III

Species	GenBank Acc#	Peptide Sequence	M.W.	Position
Desulfurobacterium thermolithotrophum	AJ001049	RAQPLSLVASG*	1097.40	1079-1136
Uncultured Aquificales OPS132	AF027104	RAQPLSCVTSG*	1117.40	1074-1131
Bacteroides caccae	X83951	RAQPLSSVTNRSC*	1417.70	1069-1126
Actinomyces bovis	X81061	RAQPLSRVASTLWWGLAGD	2083.60	1088-1145
Actinomyces meyeri	X82451	RAQPLPYVASTLWWGLVGD	2128.60	1082-1139
Denitrobacterium detoxificans	AF079507	RAQPLPHVASIRLGTHGG	1866.50	1039-1094
Uncultured GNS bacteria BPC110	AF154084	RAQPLLYVIRVIPD	1652.10	1074-1116
Uncultured GNS bacteria GCA004	AF154104	RAQPSLYVTRIIRD	1687.10	1080-1122
Uncultured GNS bacteria GCA112	AF154100	RAQPSPYVIRVIRD	1669.00	1082-1124
Acetobacter aceti	AF127399	RAQPLSLVASMFGWAL*	1746.30	1038-1095
Gluconobacter asaii	AB024492	RAQPLSLVASTFRWAL*	1815.30	1034-1092
Burkholderia sp. JB1	X92188	RAQPLSLVATQEHSRET	1922.20	1094-1144
Denitrobacter permanens	Y12639	RAQPLPLVATFSWAL*	1669.10	1077-1131
Desulfobacter curvatus	M34413	RAQPLSLVASTLCGNSNET	1960.40	1116-1172
Desulfobulbus sp. BG25	U85473	RAQPLPIVASSAGHSKGT	1863.40	1114-1170
Legionella anisa	X73394	RAQPLSLVAST*	1141.40	1078-1135
Benzene mineralizing clone SB-1	AF029039	RAQPLPLVANRSSWGL*	1764.20	1077-1134
Escherichia coli	E05133	RAQPLSFVASGPAGNSKET	1916.30	1103-1159
Uncultured Acidobacterium Sub.Div-1	Z73363	RAQPLSLVASGSSRAL*	1612.10	775-832
Uncultured Acidobacterium Sub.Div-1	Z73365	RAQPLSSVAIGSSRATLAK	1912.50	777-835
Uncultured Acidobacterium Sub.Div-1	Z73368	RAQPLFASCHH*	1933.50	779-835
Uncultured Acidobacterium Sub.Div-1	Z73364	RAQPLFAQLPSFSWALCRN	2204.80	778-835
Uncultured Acidobacterium Sub.Div-1	U68659	RAQPLLPXAII*	1218.70	573-630
Uncultured Acidobacterium Sub.Div-1	D26171	RAQPLLPVATI*	1177.50	1035-1090
Uncultured Acidobacterium Sub.Div-1	X97101	RAQPLSPVAII*	1163.50	943-998
Uncultured Acidobacterium Sub.Div-1	X97098	RAQPLSSVATI*	1141.40	1054-1109
Uncultured Acidobacterium Sub.Div-1	AF047646	RAQPLFLVATI*	1227.60	1035-1090
Uncultured Acidobacterium Sub.Div-1	AF050548	RAQPSSLVANTLW*	1441.70	572-629
Uncultured Acidobacterium Sub.Div-2	U68612	RAQPLHVATRKRELYVD	2150.60	577-630

Uncultured Acidobacterium Sub.Div-2	Y07646	RAQPLHVVATPQGGTLRG	1857.30	1085-1140
Uncultured Acidobacterium Sub.Div-3	X97097	RAQPSSLVANPQGKHPKGT	1972.40	1060-1116
Uncultured Acidobacterium Sub.Div-3	U68648	RARPLSCVAII*	1197.70	574-629
Uncultured Acidobacterium Sub.Div-3	AF013515	RAQPLSCVANPQGCTLRR	1969.50	1057-1112
Uncultured Acidobacterium Sub.Div-3	AF027004	RAQPSPCVATPPRAGALSGD	1950.40	1036-1096

* Indicates an in-frame stop codon was encountered within the polymorphic sequence.

-55-

CLAIMS

I claim:

     1. A method to determine the genetic diversity of a population comprising the steps:

        a) providing a sample of interest containing a population of interest;

        b) isolating genomic DNA from the sample;

        c) performing PCR on the genomic DNA to amplify a subregion comprising a polymorphic site of interest using an upstream primer comprising a binding site for a solid support and a downstream primer to produce an amplified fragment;

        d) cleaving the amplified fragment with a restriction enzyme that cleaves the amplified fragment at a site close to the polymorphic region;

        e) binding the cleaved amplified fragment to the solid support to form an immobilized product;

        f) splitting the immobilized product into two pools;

        g) attaching a double-stranded linker to the immobilized product of each pool, wherein the linker comprises a Type IIS restriction enzyme site;

        h) digesting the product with the Type IIS restriction enzyme to form a tag;

        i) ligating the tags to one another to form ditags;

        j) amplifying the ditags by PCR;

        k) cleaving the ditags with the restriction enzyme of step c) and ligating the ditags to form a concatemer;

        l) cloning the concatemer into a vector; and

        m) determining the sequence the concatemer.

-56-

2. The method according to claim 1, further comprising the step of digesting the tags generated in step g) with T4 DNA polymerase to create blunt ends.

3. The method according to claim 1, further comprising the step of amplifying the ditags a second time after step j).

4. The method according to claim 1, wherein the population of interest is selected from the group consisting of a microbial population, a viral population and an immune cell population.

5. The method according to claim 4, wherein the microbial population is from a sample associated with a site for petroleum or natural gas exploration.

6. The method according to claim 4, wherein the microbial population is from a sample associated with a site of oil or gas reserves.

7. The method according to claim 4, wherein the microbial population is from a sample associated with a site for mineral exploration.

8. The method according to claim 4, wherein the microbial population is from a sample associated with an agricultural field.

9. The method according to claim 4, wherein the microbial population is from a patient sample.

10. The method according to claim 9, wherein the patient is suspected to have a bacterial or fungal infection.

-57-

11. The method according to claim 4, wherein the microbial population is from a sample associated with a bioremediation site.

12. The method according to claim 4, wherein the immune cell population is from a patient sample.

13. The method according to claim 12, wherein the patient is suffering from a disease selected from the group consisting of autoimmune disease, graft v. host disease, lymphoma, leukemia, Hodgkin's disease, cancer, atherosclerosis and psoriasis.

14. The method according to claim 4, wherein the microbial population is from a sample of an insect or a parasite.

15. A method for analyzing a marker diversity profile (MDP) by comparison to a marker diversity profile matrix database to identify at least one parameter associated with the MDP, comprising the steps of:

a) constructing a marker diversity profile comprising a plurality of nucleic acid markers,

wherein each marker is an identifier for one member or a limited number of members of a population,

wherein the abundance of said marker provides an indication as to the abundance of the corresponding member of said population, and

wherein said population is associated with at least one parameter;

b) detecting the abundance of each marker;

c) transducing the abundance of each marker into an electrical signal;

d) storing a marker diversity profile comprising the marker electrical signals in a marker diversity profile matrix data structure;

e) determining at least one parameter associated with said marker diversity profile by comparing said marker diversity profile with a

-58-

marker diversity profile matrix data structure comprising marker diversity profiles that are each known to be associated with at least one parameter.

16. The method according to claim 15, wherein the population is a microbial population.

17. The method according to claim 15, wherein a marker diversity profile is compared to a marker diversity profile matrix data structure according to comparison rules contained in a knowledge base.

18. The method according to claim 15, wherein the abundance of markers in said marker diversity profile is expressed as the relative abundance in the population.

19. The method according to claim 15, wherein the markers of said marker diversity profile are detected through DNA sequencing.

20. The method according to claim 15, wherein the markers of said marker diversity profile are detected by determining the masses of the corresponding translation products.

21. The method according to claim 15, wherein the markers of said marker diversity profile are detected through hybridization.

22. The method according to claim 15, wherein the markers of said marker diversity profile are detected through size differences of the markers.

-59-

23. The method according to claim 22, wherein the size differences of the markers correlates to a size difference of a 16S and 23S rDNA intergenic region.

24. The method according to claim 22, wherein the size differences of the markers correlates to a size difference of a 23S and 5S rDNA intergenic region.

25. The method according to claim 22, wherein the size differences of the markers are identified by denaturing gradient gel electrophoresis (DGGE) or terminal restriction fragment length polymorphism (T-RFLP).

26. The method according to claim 15, wherein said population is comprised of immunoglobulin genes.

27. The method according to claim 15, wherein said population is comprised of viral, viroid or bacteriophage genes.

28. A method for analyzing a marker diversity profile database to identify one or more robust indicator markers for a desired parameter, comprising the steps of:
   a) constructing a marker diversity profile comprising a plurality of nucleic acid markers,

   wherein each marker is an identifier for one member or a limited number of members of a population,

   wherein the presence or abundance of said marker provides an indication as to the presence or abundance of the corresponding member of said population, and

   wherein said population is associated with said desired parameter;
   b) detecting the presence or abundance of each marker;

-60-

c) transducing the presence or abundance of each marker into an
electrical signal;

d) storing the marker diversity profile comprising marker electrical
signals in a marker diversity profile matrix data structure;

e) correlating the presence or abundance of each marker in the marker
diversity profile database with the presence or abundance of each said
parameters;

f) ranking the markers for each parameter based upon correlation
values; and

g) designating one or more markers as robust indicators for the
parameter as those that have a correlation value of at least 0.4 to the
parameter.

29. The method according to claim 28, wherein the correlation value is
at least 0.5.

30. The method according to claim 28, wherein the correlation value is
at least 0.6.

31. The method according to claim 28, wherein the correlation value is
at least 0.7.

32. The method according to claim 28, wherein the correlation value is
at least 0.8.

33. The method according to claim 28, wherein the correlation value is
at least 0.9.

-61-

34. The method according to claim 28, wherein the correlation value is at least 0.95.

35. The method according to claim 28, wherein the presence of said marker correlates with either the presence or the abundance of said parameter.

36. The method according to claim 28, wherein the abundance of said marker correlates with either the presence or abundance of said parameter.

37. A method for analyzing a MDP database to identify groups of correlated markers, comprising the steps of:

a) constructing a marker diversity profile comprising a plurality of nucleic acid markers,

wherein each marker is an identifier for one member or a limited number of members of a population,

wherein the presence or abundance of said marker provides an indication as to the presence or abundance of a corresponding member of said group, and

wherein said group is associated with a desired parameter;

b) detecting the presence or abundance of each marker;

c) transducing the presence or abundance of each marker into an electrical signal;

d) storing the marker diversity profile comprising marker electrical signals in a marker diversity profile matrix data structure;

e) correlating the presence or abundance of each marker in the marker diversity profile database with the presence or abundance of other markers in pairwise comparison to form a correlation matrix;

f) separating the markers into correlated groups based upon values from correlation matrix.

-62-

38. A method of constructing a matrix diversity profile database, comprising the steps of

a) constructing a marker diversity profile comprising a plurality of nucleic acid markers,

wherein each marker is an identifier for one member or a limited number of members of a population,

wherein the presence or abundance of said marker provides an indication as to the presence or abundance of the corresponding member of said population, and

wherein said population is associated with at least one parameter;

b) detecting the presence or abundance of each marker;

c) transducing the presence or abundance of each marker into an electrical signal;

d) storing the marker diversity profile comprising marker electrical signals in a marker diversity profile matrix data structure; and

e) indexing the marker diversity profile with the associated parameter to form a matrix diversity profile database.

39. A method for identifying a marker that correlates with a desired parameter, comprising the steps of

a) constructing a marker diversity profile comprising a plurality of nucleic acid markers,

wherein each marker is an identifier for one member or a limited number of members of a population,

wherein the presence or abundance of said marker provides an indication as to the presence or abundance of the corresponding member of said population, and

wherein said population is associated with said desired parameter;

b) detecting the presence or abundance of each marker;

-63-

c) correlating the presence or abundance of each marker in the marker diversity profile database with the parameter;

d) ranking the markers for the parameter based upon correlation values; and

e) designating one or more markers as a robust indicator for the parameter as those that have a correlation value of at least 0.4 to the parameter.

40. The method according to claim 39, wherein said desired parameter is the presence of mineral deposits and/or petroleum reserves.

41. A method for identifying a marker that correlates with a desired parameter, comprising the steps of

a) constructing a marker diversity profile comprising a plurality of nucleic acid markers,

wherein each marker is an identifier for one member or a limited number of members of a population,

wherein the presence or abundance of said marker provides an indication as to the presence or abundance of the corresponding member of said population, and

wherein said population is associated with said desired parameter;

b) detecting the presence or abundance of each marker;

c) correlating the presence or abundance of each marker in the marker diversity profile database with the parameter;

d) ranking the markers for the parameter based upon correlation values;

e) designating one or more markers as a potential indicator for the parameter as those that have a correlation value of at least 0.4 to the parameter; and

-64-

f) determining whether the potential indicator is present or abundant in a population that is not associated with said desired parameter; and

g) designating one or more markers as a robust indicator for the parameter if the potential indicator is not present or is less abundant in a population that is not associated with said desired parameter than in a population that is associated with said desired parameter.

42. The method according to claim 41, wherein said desired parameter is the presence of mineral deposits and/or petroleum reserves.

43. A method for performing mineral deposit exploration or subsurface petroleum or natural gas reserve exploration, comprising the steps of:

a) isolating nucleic acids from an environmental sample;

b) amplifying a specific DNA molecule that has been shown to correlate with the presence of mineral deposits and/or petroleum or natural gas reserves;

c) detecting the presence of said amplified DNA molecule; and

c) inferring the presence of said mineral deposits and/or said petroleum or natural gas reserves from the presence of said amplified DNA molecule.

44. A method for performing mineral deposit exploration or subsurface petroleum or natural gas reserve exploration, comprising the steps of:

a) isolating nucleic acids from an environmental sample;

b) amplifying a specific RNA molecule that has been shown to correlate with the presence of mineral deposits and/or petroleum or natural gas reserves;

c) detecting the presence of said amplified RNA molecule; and

c) inferring the presence of said mineral deposits and/or said petroleum or natural gas reserves from the presence of said amplified RNA molecule.

Bacteria
16s rRNA Polymorphic Regions

Figure 1.

16s rRNA Polymorphic Regions

STEP

I

II

III

IV

V

VI
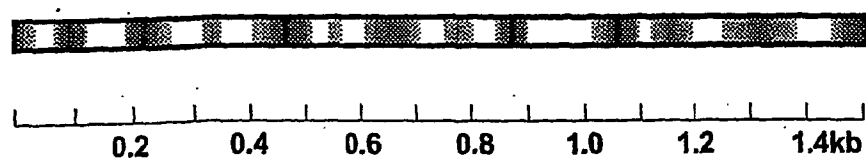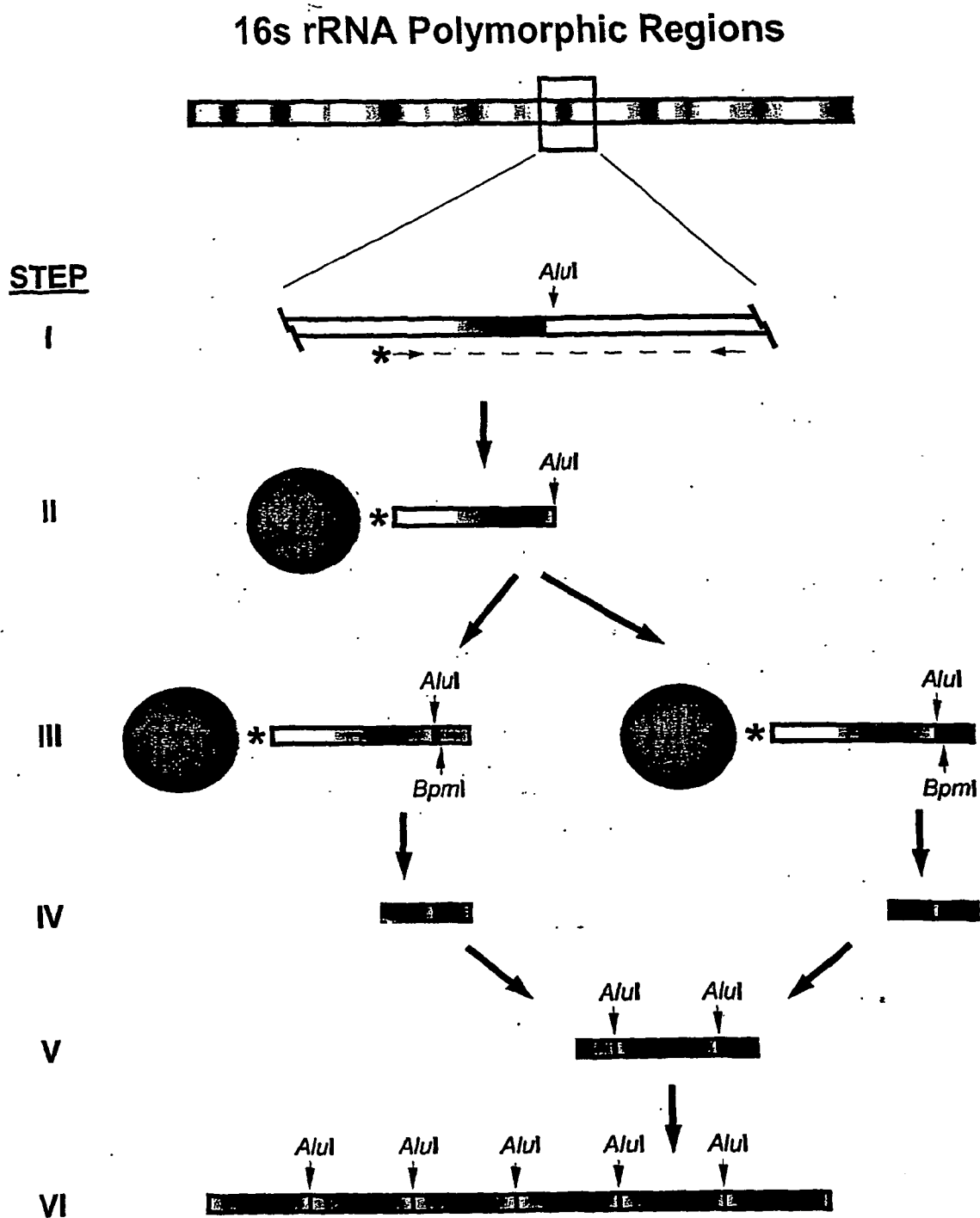
Figure 2.

Bacteria

Acidobacteria Group
          Various uncultured environmental Acidobacteria
Aquificales
    Desulfurobacterium Group
        Desulfurobacterium
            Desulfurobacterium thermolithotrophum
      environmental samples
           unidentified Aquificales OPS132
CFB/Green sulfur bacteria group
      Bacteroidaceae
        Bacteroides
          Bacteroides caccae
Firmicutes(gram-positive bacteria)
    Actinobacteria(high G+C gram-positive bacteria)
        Actinomycetaceae
          Actinomyces
            Actinomyces bovis
            Actinomyces meyeri
          Denitrobacterium
           Denitrobacterium detoxificans
Green non-sulfur bacteria
          environmental samples
            uncultured HC-seep bacterium BPC110
            uncultured HC-seep bacterium GCA004
            uncultured HC-seep bacterium GCA112
Proteobacteria(purple non-sulfur bacteria)
    alpha subdivision
        Acetobacteraceae
         Acetobacter
           Acetobacter aceti
        Gluconobacter
          Gluconobacter asaii
    beta subdivision
      Burkholderia group
        Burkholderia
          Burkholderia **sp. JB1**
        Denitrobacter
          Denitrobacter permanens
    delta subdivision
        Desulfobacter
          Desulfobacter curvatus
        Desulfobulbus
          Desulfobulbus **sp. BG25**
      Legionellaceae
        Legionella
          Legionella anisa
unclassified Bacteria
          benzene mineralizing clone SB-1
          uncultured eubacterium env.OPS 1

Fig. 3

Archaea (archaebacteria)

Crenarchaeota (extremely thermophilic archaebacteria)
Desulfurococcales
    Desulfurococcaceae
        Aeropyrum
            Aeropyrum pernix
        Desulfurococcus
            Desulfurococcus mobilis
        Staphylothermus
            Staphylothermus marinus
Sulfolobales
    Metallosphaera
        Metallosphaera sedula
    Sulfolobus
        Sulfolobus acidocaldarius
        Sulfolobus metallicus
Thermoproteales
    Thermoproteaceae
        Caldivirga
            Caldivirga maquilingensis
    Pyrobaculum
        Pyrobaculum islandicum

Euryarchaeota
Archaeoglobales
    Archaeoglobaceae
        Archaeoglobus
            Archaeoglobus fulgidus
            Archaeoglobus veneficus
Halobacteriales
    Halobacteriaceae
        Haloarcula
            Haloarcula japonica
        Halococcus
            Halococcus morrhuae
Methanococcales
    Methanococcaceae
        Methanococcus
            Methanococcus jannaschii
Methanobacteriales
    Methanobacteriaceae
        Methanobacterium
            Methanobacterium bryantii
            Methanobacterium subterraneum
Thermococcales
    Thermococcaceae
        Pyrococcus
            Pyrococcus abyssi
Thermoplasmales
    Picrophilaceae
        Picrophilus
            Picrophilus oshimae

Fig. 4

Fig. 5

Fig. 6

A

Positive Correlation

r = 0.90

Relative Abundance of Each Marker

Relative Abundance of
Each Marker

B

Zero Correlation with All Markers

r = 0.10

Relative Abundance of Each Marker

Relative Abundance of
Each Marker

C

Positive Correlation with Subgroup of Markers

r = 0.95

Relative Abundance of Each Marker

Relative Abundance of
Each Marker

Fig. 7

Generating Marker
Diversity/Constituent Profile
Database

Set N=1 — 100

Assemble Polymorphic
Markers - N — 105

Detect Polymorphic
Markers — 110

Transduce each marker
into electrical output
signal — 115

140 — Increment N

Detect parameters — 120

Transduce each parameter
into electrical output
signal — 125

Store output signals
into corresponding output
matrix data structure — 130

No ◇ Done forming
data structures? — 135

Yes

diversity

Form marker ~~constituent~~
profile database from
N data structures — 145

Figure 8.

Diversity

Marker ~~Constituent~~ Profile — 200

MDP

Compare to ~~MCP~~ Database — 205

215 — Same Parameters

Yes

MDP New ~~MCP~~ = Database ~~MCP~~ *MDP* — 210

No

MDP New ~~MCP~~ is subset of Database ~~MCP~~ *MDP* — 220

No — Parameters undefined — 225

Yes

245 — Parameters undefined

Marker-Marker correlation analysis between common subset of markers — 230

No

240 — Markers and parameter correlated

Yes

All common markers correlated — 235

Yes

250 — Parameter Identified

No

Sort common markers into correlated groups — 255

260 — Correlated Group N

Correlated Group N + 1 — 280

265 — Marker-Parameter Correlated

270 — Parameters undefined

No

285 — Marker-Parameter Correlated

No — Parameters undefined — 290

Yes

275 — Parameter identified

Yes — Parameter identified — 295

Figure 9.

Figure 10.

Figure 11.

SARD OLIGONUCLEOTIDES

```
TX-003    5'-    Phosphate-GCTCCAGGTCTACATCCTAGTCAGGACddC-3'
TX-012    5'-    Phosphate- CTCCAGGTCTACATCCTAGTCAGGACddC
TX-010R   3'              GAGGTCCAGATGTAGGATCAGTCCTGGATA-5'
TX-013    3'              GAGGTCCAGATGTAGGATCAGTCCTGGATA-5'
TX-004R   3'                  CAGATGTAGGATCAGTCCTGGATA-5'
TX-007R   3'-                 CAGATGTAGGATCAGTCCTGG -Biotin-5'
TX-111    3'-                 CAGATGTAGGATCAGTCCTGGCACTGATAGCTC -Biotin-5'

TX-005    5'-    Phosphate-GCTCCAGACTAGCATCCGCTGACTTGAddC
TX-014    5'-    Phosphate- CTCCAGACTAGCATCCGCTGACTTGAddC
TX-011R   3'              GAGGTCTGATCGTAGGCGACTGAACTGTAA
TX-015    3'              GAGGTCTGATCGTAGGCGACTGAACTGTAA
TX-006R   3'                  TGATCGTAGGCGACTGAACTGTAA
TX-008R   3'                  TGATCGTAGGCGACTGAACTG -Biotin-5'
TX-121    3'                  TGATCGTAGGCGACTGAACTGGCAACGTTGGAC -Biotin-5'
```

```
TX-001    5'-    Biotin-GTG TAG HRG TGA AAT DCD YA

TX-002    5'-    YTC ACG RCA YGA GCT GAC GAC

TX-003    5'-    Phosphate-GCT CCA GGT CTA CAT CCT AGT CAG GACddC

TX-004    5'-    ATA GGT CCT GAC TAG GAT GTA GAC

TX-005    5'-    Phosphate-GCT CCA GAC TAG CAT CCG CTG ACT TGAddC

TX-006    5'-    AAT GTC AAG TCA GCG GAT GCT AGT

TX-007    5'-    Biotin-GGT CCT GAC TAG GAT GTA GAC

TX-008    5'-    Biotin-GTC AAG TCA GCG GAT GCT AGT

TX-009    5'-    Biotin-GGA TTA GAW ACC CBG GTA GTC

TX-010    5'-    ATA GGT CCT GAC TAG GAT GTA GAC CTG GAG

TX-011    5'     AAT GTC AAG TCA GCG GAT GCT AGT CTG GAG

TX-012    5'-    Phosphate-CTC CAG GTC TAC ATC CTA GTC AGG ACddC

TX-013    5'-    ATA GGT CCT GAC TAG GAT GTA GAC CTG GAG

TX-014    5'-    Phosphate-CTC CAG ACT AGC ATC CGC TGA CTT GAddC

TX-015    5'     AAT GTC AAG TCA GCG GAT GCT AGT CTG GAG

TX-111    5'     Biotin-CTC GAT AGT CAC GGT CCT GAC TAG GAT GTA GAC

TX-121    5'     Biotin-CAG GTT GCA ACG TCA AAG TCA GCG GAT GCT AGT
```

Fig. 12

SARD STRATEGY

1. Amplify rDNA with TX-009 and 1392R

```
Bio---------nnnnNNNNNNNNNNNNNAGCTnnnnnnnnnnnnnnnnnn
        ---------nnnnNNNNNNNNNNNNNNTCGAnnnnnnnnnnnnnnnnnn
```

2. Cut with *AluI*.

```
Bio---------nnnnNNNNNNNNNNNNNAG        CTnnnnnnnnnnnnnnnnnn
        ---------nnnnNNNNNNNNNNNNNNTC        GAnnnnnnnnnnnnnnnnnn
```

3. Split into two pools, bind to SA beads, wash, add TX-12/13 adapter or TX-14/15 adapter.

```
Bio-----------nnnnNNNNNNNNNNNNNAG        CTCCAGGTCTACATCCTAGTCAGGACddC -3'   (TX012)
        ---------nnnnNNNNNNNNNNNNNNTC        GAGGTCCAGATGTAGGATCAGTCCTGGATA-5'   (TX013)  .
```

                    OR

```
Bio----------nnnnNNNNNNNNNNNNNAG        CTCCAGACTAGCATCCGCTGACTTGAddC -3'   (TX014)
        --------nnnnNNNNNNNNNNNNNNTC        GAGGTCTGATCGTAGGCGACTGAACTGTAA-5'   (TX015)
```

4. Ligate adapters, wash.

```
Bio---------nnnnNNNNNNNNNNNNNNAGCTCCAGGTCTACATCCTAGTCAGGACddC -3'   (TX012)
        --------nnnnNNNNNNNNNNNNNNTCGAGGTCCAGATGTAGGATCAGTCCTGGATA-5'   (TX013)
```

5. Cleave with *BpmI*, remove SA-beads.

```
        NNNNNNNNNNNNNNNNAGCTCCAGGTCTACATCCTAGTCAGGACddC -3'   (TX012)
        nnNNNNNNNNNNNNNNTCGAGGTCCAGATGTAGGATCAGTCCTGGATA-5'   (TX013)
```

6. Remove 3' termini with PolI plus 4 dNTPs, combine pools, ligate.

```
(TX015)                                                                          (TX012)
AATGTCAAGTCAGCGGATGCTAGTCTGGAGCTNNNNNNNNNNNNNNNNNNNNNNNNNNAGCTCCAGGTCTACATCCTAGTCAGGACddC -3'
 ddCAGTTCAGTCGCCTACGATCAGACCTCGANNNNNNNNNNNNNNNNNNNNNNNNNNNNTCGAGGTCCAGATGTAGGATCAGTCCTGGATA-5'
(TX014)                                                                          (TX013)
```

7. Amplify ditags with TX-111, TX-121 primers

```
BioCAGGTTGCAACGGTCAAGTCAGCGGATGCTAGT---------->
        AATGTCAAGTCAGCGGATGCTAGTCTGGAGCTNNNNNNNNNNNNNNNNNNNNNNNNAGCTCCAGGTCTACATCCTAGTCAGGACddC -3'
        ddCAGTTCAGTCGCCTACGATCAGACCTCGANNNNNNNNNNNNNNNNNNNNNNNNNNNTCGAGGTCCAGATGTAGGATCAGTCCTGGATA-5'
                                                        <-------CAGATGTAGGATCAGTCCTGGCACTGATAGCTC-Bio
```

8. Purify PCR product, cleave with *AluI*

```
BioCAGGTTGCAACGGTCAAGTCAGCGGATGCTAGTCTGGAG  CTNNNNNNNNNNNNNNNNNNNNNNNNAG   CTCCAGGTCTACATCCTAGTCAGGACCGTGACTATCGAG-3'
        GTCCAACGTTGCCAGTTCAGTCGCCTACGATCAGACCTC   GANNNNNNNNNNNNNNNNNNNNNNNTC   GAGGTCCAGATGTAGGATCAGTCCTGGCACTGATAGCTC-Bio
```

9. Purify away adapters, ligate.

```
CTNNNNNNNNNNNNNNNNNNNNNNNNNNAGCTNNNNNNNNNNNNNNNNNNNNNNNNNNAGCTNNNNNNNNNNNNNNNNNNNNNNNNNNAG
GANNNNNNNNNNNNNNNNNNNNNNNTCGANNNNNNNNNNNNNNNNNNNNNNNTCGANNNNNNNNNNNNNNNNNNNNNNNNNTC
```

10. Gel-purify large concatemers, clone into *SmaI* site of vector.

Fig. 13

Fig. 14

ACGATGAGCACTAGCT
(1)

ACGATGAGTACTAGCT
(1)

ACGATGATGACTAGCT
(2)

ACGATGGATGCTAGCT
(1)

ATGCTAGTCTGGAGCT
(6)

ATGGCTGTCGTCAGCT
(50)

ATGGTTGTCGTCAGCT
(1)

ATTCCGTGCCGTAGCT
(1)

CACTAGTGGCGCAGCT
(2)

CCCCCGTGCCGAAGCT
(1)

CCCCCGTGCCGCAGCT
(1)

CCCCCTTCCTCCAGCT
(1)

CCCCGGTGCCGCAGCT
(1)

CCGGGTAGTCCCAGCT
(5)

CCTCCGTGCCGAAGCT
(3)

CCTCCGTGCCGCAGCT
(2)

CCTCCGTGCTGCAGCT
(1)

CCTCGGCGCCGCAGCT
(2)

CCTCGGTGCCGCAGCT
(1)

CCTCGGTGTCGCAGCT
(2)

CCTGGGTGCCGCAGCT
(2)

CCTGTGTGACGAAGCT
(1)

CCTTGGTAACGAAGCT
(1)

CCTTGGTACCGAAGCT
(1)

CGCCAGTGCCGTAGCT
(1)

CGCCTGTGCCGTAGCT
(2)

CGTCCGTGCCGAAGCT
(2)

CGTCCGTGCCGCAGCT
(1)

CGTCGGTGCCGCAGCT
(3)

CTCCCGTGCCGCAGCT
(1)

CTCCCGTGCCGGAGCT
(1)

CTCCGGTGCCGCAGCT
(1)

CTCCTGTGCCGAAGCT
(1)

CTCCTGTGCCGCAGCT
(2)

CTGCCGTGCCGAAGCT
(9)

CTGCTGTGCCGAAGCT
(2)

CTGTCGTGCCGAAGCT
(1)

CTTCAGTATCGAAGCT
(1)

CTTCCGCGCCGGAGCT
(2)

CTTCCGTGCCGCAGCT
(3)

CTTCCGTGCCGGAGCT
(1)

CTTCGGTGCCGCAGCT
(1)

CTTCTGTGGCGAAGCT
(1)

GATCCGTGCCGTAGCT
(1)

GCTCTGTGCCGAAGCT
(1)

GCTGGGTGCCCAAGCT
(1)

GGTCCGTGCCGCAGCT
(1)

GGTGCTCTTCGGAGCT
(2)

GTAAACGATGGAAGCT
(1)

GTGGCTGTCGTCAGCT
(2)

GTTCCGTGCCGAAGCT
(2)

GTTCCGTGCCGCAGCT
(3)

TATCAGTGGCGCAGCT
(1)

TCTCCGTGCCGCAGCT
(1)

TCTCTGTGCCGCAGCT
(2)

TCTCTGTGCCGTAGCT
(1)

TGTCCGTGCCGTAGCT
(1)

TTTCCGTGCCGCAGCT
(2)

Fig. 15

ACGATGATAACTAGCT   CCGGGTAGTCCCAGCT   CTCCAGTGCCGCAGCT   GATCCGTGCCGCAGCT
(2)   (4)   (1)   (2)

ACGATGATGACTAGCT   CCGGGTAGTCCTAGCT   CTCCCGTGCCACAGCT   GCTCCGTGCCGAAGCT
(1)   (1)   (1)   (1)

ACGATGGGCACTAGCT   CCTCCGTGCCGAAGCT   CTCCCGTGCCGCAGCT   GCTCCGTGCCGTAGCT
(1)   (2)   (2)   (1)

ACGGCTGTCGTCAGCT   CCTCCGTGCCGCAGCT   CTCCCGTGCCGGAGCT   GCTCTGTGCCGAAGCT
(2)   (2)   (3)   (2)

ACTACGAGCGCAAGCT   CCTCCGTGCCGGAGCT   CTCCGGTGCCGCAGCT   GCTCTGTGCCGTAGCT
(1)   (1)   (1)   (1)

ACTTAATGCGTTAGCT   CCTCGTAAGGGGAGCT   CTCCTGTGCCGCAGCT   GGGCTTGTCGTCAGCT
(1)   (1)   (1)   (1)

ATGCTAGTCTGGAGCT   CCTGGGTGCCGCAGCT   CTGCCGCGCCGGAGCT   GTAAACGATGGAAGCT
(8)   (1)   (1)   (1)

ATGGCTCTCGTCAGCT   CCTGGTAGTCCCAGCT   CTGCCGTGCCGAAGCT   GTTCCGTGCCGCAGCT
(1)   (2)   (5)   (2)

ATGGCTGTCGCCAGCT   CCTGGTAGTCCTAGCT   CTGCCGTGCCTAAGCT   GTTCCGTGCCGTAGCT
(3)   (1)   (1)   (2)

ATGGCTGTCGTCAGCT   CCTTAGTAACGCAGCT   CTGCGGTGCCGCAGCT   GTTCTGTGCCGCAGCT
(70)   (1)   (1)   (2)

ATGGTTGTCGTCAGCT   CCTTGGTAACGAAGCT   CTGCTGTGCCGAAGCT   TCTCACGACACGAGCT
(2)   (1)   (7)   (2)

ATGTAGACCTGGAGCT   CGCCAGTGCCGAAGCT   CTGTCGTGCCGAAGCT   TCTCAGTAACGTAGCT
(12)   (1)   (1)   (1)

ATTCCGTGCCGCAGCT   CGCCGGTGCCGCAGCT   CTTCCGCGCCGGAGCT   TCTCCGTGCCGCAGCT
(6)   (2)   (7)   (2)

ATTCCGTGCCGTAGCT   CGCCTGTGCCGTAGCT   CTTCCGTGCCGAAGCT   TCTCTGTGCCGCAGCT
(1)   (2)   (2)   (4)

CACAAGCGGTGGAGCT   CGCTCGTGGCGAAGCT   CTTCCGTGCCGCAGCT   TGGACGTTGCGGAGCT
(1)   (1)   (6)   (2)

CACTAGTGGCGCAGCT   CGGAGGCGTCGTAGCT   CTTCCGTGCCGGAGCT   TTTCCGTGCCGGAGCT
(4)   (1)   (2)   (1)

CATCCGTGCCGAAGCT   CGTCAGTGTCGCAGCT   CTTCGGTGCCGCAGCT
(1)   (2)   (1)

CCCCAGGGCCCAAGCT   CGTCCGTGCCGAAGCT   CTTCGGTGTCGCAGCT
(1)   (7)   (1)

CCCCGGTGCCGCAGCT   CGTCCGTGCCGCAGCT   CTTGGGTGCCGCAGCT
(1)   (1)   (2)

CCCGCGTGCCGGAGCT   CGTCCGTGCCGGAGCT   CTTTAGTAACGCAGCT
(1)   (1)   (2)

CCGCGGTGCCGTAGCT   CGTCGGTGCCGCAGCT   GACCCGCAAGGGAGCT
(1)   (2)   (1)

Fig. 16

Wy-1 SARD Tag Diversity Profile

Wy-2 SARD Tag Diversity Profile

Fig. 17